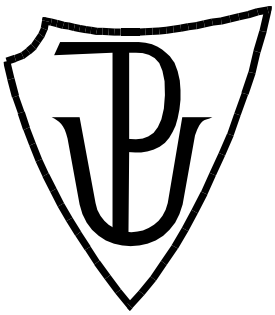
UNIVERZITA PALACKÉHO V OLOMOUCI

Přírodovědecká fakulta

Katedra biochemie



Analýza repetitivních sekvencí B chromozomu u druhu *Sorghum purpureosericeum*

DIPLOMOVÁ PRÁCE

Autor:		**Karel Vrabka**
Studijní program:		N1406 Biochemie
Studijní obor:		Bioinformatika
Forma studia:		Prezenční
Vedoucí práce:		**Mgr. Jan Bartoš, Ph.D.**
Rok:	:	2021

Prohlašuji, že jsem diplomovou práci vypracoval/a samostatně s vyznačením všech použitých pramenů a spoluautorství. Souhlasím se zveřejněním diplomové práce podle zákona č. 111/1998 Sb., o vysokých školách, ve znění pozdějších předpisů. Byl/a jsem seznámen/a s tím, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorský zákon, ve znění pozdějších předpisů.

V Olomouci dne

Poděkování

Tímto způsobem bych rád poděkoval Mgr. Janu Bartošovi, PhD za odborné vedení diplomové práce, vstřícnost, trpělivost a cenné rady. Rovněž nemohu opomenout poděkovat své rodině, kamarádům a přítelkyni Žanetě, bez jejichž podpory by tato práce nemohla vzniknout.

Bibliografická identifikace

Jméno a příjmení autora	Bc. Karel Vrabka
Název práce	Analýza repetitivních sekvencí B chromozomu u druhu *Sorghum purpureosericeum*
Typ práce	Diplomová
Pracoviště	Katedra biochemie
Vedoucí práce	Mgr. Jan Bartoš, Ph.D.
Rok obhajoby práce	2021
Abstrakt	B chromozomy jsou nadpočetnou součástí karyotypu podléhající specifickému způsobu dědičnosti, které se vyskytují v genomech rozličných druhů rostlin i zvířat. Jejich přítomnost byla odhalena rovněž u rodu *Sorghum*, jehož zástupci mají hospodářský význam a i proto je podstatná správná identifikace genotypů jednotlivých rostlin. V této práci byla provedena úprava sekvenačních dat z rodu *Sorghum purpureosericeum*, na datech byla poté provedena komparativní klastrová analýza softwarem *RepeatExplorer*, jež vedla k identifikaci repetitivních sekvencí v genomu. Součástí výsledků práce je rovněž vyvinutý software ke zpracování a vizualizaci výsledků komparativní analýzy.
Klíčová slova	B chromozomy, *Sorgum purpureosericeum*, RepeatExplorer, repetitivní sekvence DNA, DNA
Počet stran	55
Počet příloh	6
Jazyk	Český

Bibliographical identification

Autor's first name and surname	Karel Vrabka
Title	Analysis of repetitive DNA sequences in genome of *Sorghum Purpureosericeum*
Type of thesis	Diploma
Department	Department of biochemistry
Supervisor	Mgr. Jan Bartoš, Ph.D.
Abstract	B chromosomes are supernumerary part of karyotypes characterized by specific type of chromosomes transmission rates, which are part of genomes of various plants and animals. Presence of B chromosomes was also discovered in species of *Sorghum* genus. *Sorghum* species are not only used in food industry, but also represent important source of genes, that could be used in breeding. This thesis describes a workflow for comparative analysis of repeat DNA from *Sorghum purpureosericeum* using *RepeatExplorer* cluster analyses. The software implemented for processing processing and visualization of *RepatExplorer* output files is a part of thesis result.
The year of presentation	2021
Keywords	B chromosomes, *Sorgum purpureosericeum*, RepeatExplorer, repetitive DNA sequences, DNA
Number of pages	55
Number of appendices	6
Language	Czech

OBSAH

CÍLE PRÁCE

1. Vypracování literární rešerše na téma B chromozomů u *S. purpureosericeum* a analýzy repetitivních sekvencí.

2. Identifikace repetitivních sekvencí specifických pro B chromozom *S. purpureosericeum* nebo na tomto chromozomu obohacených v porovnání s A chromozomy.

3. Navržení primerů pro identifikaci B chromozomů pomocí PCR.

1 ÚVOD

B chromozomy patří mezi postradatelnou část karyotypu, pro kterou je typický odlišný způsob dědičnosti, tzv. drajv. Důsledkem chromozomového drajvu je akumulace B chromozomů v genomu, která může mít jak negativní, tak pozitivní vliv na celý organismus. Ačkoli výzkum B chromozomů probíhá intenzivněji až v posledních několika desítkách let a je zaměřen především na modelové organismy a hospodářsky významné rostliny, předpokládá se, že jsou B chromozomy zastoupeny přibližně v 15 % eukaryotických organismů (Beukeboom, 1994), z nichž většinu tvoří rostliny (Jones et al., 2008). Přítomnost B chromozomů byla rovněž potvrzena u několika druhů rostlin rodu *Sorghum* (Wu, 1992).

Rod *Sorghum* zahrnuje 25 druhů rostlin, z nichž jsou nejvíce pěstovány odrůdy *Sorghum bicolor*, jež nacházejí využití především v lihovarnickém průmyslu, či jako krmivo pro zvířata, ale jsou také hlavní složkou potravy v rozvojových se zemích. Zástupci rodu *Sorghum* se vyznačují velkou tolerancí k suchu a nízkými nároky na pěstování, což jsou vlastnosti, které jsou kvůli změnám klimatu velmi žádoucí. V budoucnosti se jejich hospodářský význam s velkou pravděpodobností zvětší, což je podmíněno i zvýšením výnosu a zlepšením jiných vlastností prostřednictvím šlechtění. Planě rostoucí druhy rostlin, mezi které patří i *Sorghum purpureosericeum*, představují cenný zdroj genů, které mohou reprezentovat zdroj žádoucích znaků při šlechtění. V souvislosti s možnou přítomností B chromozomů v genomu *Sorghum purpureosericeum*, je pro jejich lokalizaci potřebné mít sondy odvozené od B-specifických sekvencí.

Pro strukturu B chromozomů jsou typické repetitivní sekvence DNA, které často mají původ v chromozomech stejného, či příbuzného druhu, ale rovněž mohou být specifické pro daný B chromozom (Perfectti a Werren, 2001). Existuje několik přístupů k analýze repetitivních sekvencí, které se odvíjí především od typu genomických dat. V případě analýzy krátkých čtení vzniklých sekvencováním technologií Illumina, lze k analýze repetitivní DNA využít software RepeatExplorer (Novák et al., 2010), jehož algoritmus je založen na klastrování čtení s nízkým pokrytím genomu na základě jejich vzájemné podobnosti. Výsledkem algoritmu jsou klastry představující jednotlivé rodiny repetitivních sekvencí v genomu. Z konsenzuální sekvence klastrů specifických pro B chromozomy lze odvodit sondy pro jejich lokalizaci.

2 SOUČASNÝ STAV ŘEŠENÉ PROBLEMATIKY

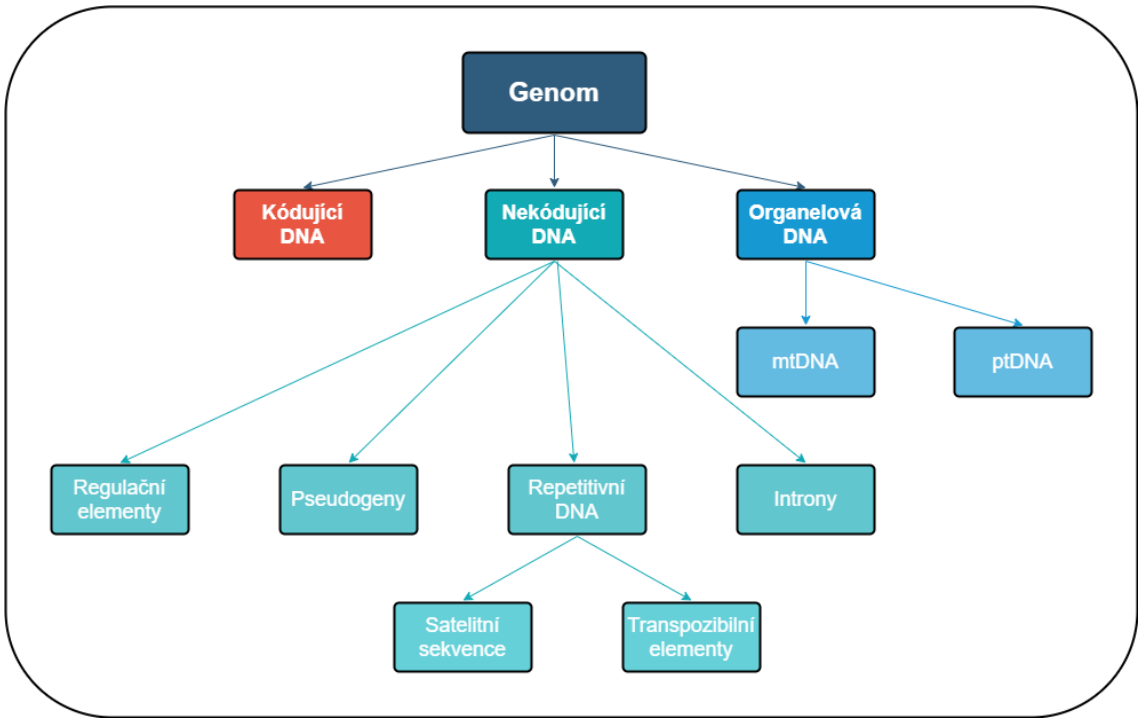
2.1 Složení genomu

Veškerá dědičná informace organismů je uložena v sekvencích nukleových kyselin, jedná se ve většině případů o DNA (deoxyribonukleová kyselina), avšak v určitých typech virů se může jednat o RNA (ribonukleová kyselina). Genomem je nazývána veškerá haploidní DNA (popř. RNA) organismu, skládající se z kódujících, nekódujících úseků (Juergen, 2009) a DNA semiautonomních organel u eukaryotických organismů (Costa, 2008).

Struktura genomu se odvíjí od složitosti organismu. Genom prokaryotických organismů je v zásadě menší než eukaryotické genomy, je rovněž průměrně z 20 % tvořen nekódujícími sekvencemi (Costa, 2008) a neobsahuje introny. Prokaryotické buňky jsou zpravidla haploidní, přičemž genetická informace je uložena v cytoplazmě ve formě jediného nukleoidu – dvoušroubovicové kruhové molekula DNA. Nicméně existují i jedinci, kteří mají genom rozdělen do více chromozomů, jenž mohou být i lineární (Brown, 2002). Mimo nukleoid je DNA v prokaryotických buňkách lokalizována v menších kruhových strukturách, plazmidech. Těch se může v buňce nacházet až několik stovek a je pro ně příznačný nerovnoměrný přenos do nově vzniklých buněk. Rovněž jsou náchylnější k sekvenčním mutacím a mnohdy poskytují jedincům evoluční výhody, např. rezistenci k antibiotikům (Fuerst, 2010). Typickým prvkem prokaryotního genomu jsou také operony – skupiny po sobě jdoucích genů, transkribovány z jednoho promoteru a také společně regulovány.

Pro eukaryotní genom je typické rozdělení DNA do dvou a více lineárních chromozomů v jádře a do zpravidla cirkulárních molekul DNA v mitochondriích a u rostlin v plastidech. DNA je v jádře uchovávána v podobě komplexu DNA a histonů (malé bazické nukleoproteiny) – chromatinu. Rozdílná forma kondenzace chromatinu má za následek odlišnou transkripční aktivitu řetězců DNA. Ve volněji uspořádaném chromatinu – euchromatinu, se nachází většina genů, oproti tomu ve více kondenzované části chromatinu – heterochromatinu se nacházejí především repetitivní a jiné nekódující sekvence DNA. Chromozómy jsou nejkondenzovanější podobou chromatinu, kterou DNA nabývá pouze v průběhu buněčného dělení.

Množství DNA v haploidním genomu konkrétního druhu se označuje jako hodnota C a u eukaryotických organismů se velmi liší, přestože se předpokládá, že se počet



Obr. 1 Klasifikace genomické DNA

kódujících genů u odlišných eukaryotických organismů výrazně neliší. Zatímco lidský genom má přibližně 3 Gb, tak genom měňavky *Amoeba dubia* obsahuje až 670 Gb. S takto vysokým rozpětím souvisí i tzv. paradox hodnoty C – velikost genomu není přímo úměrná počtu genů, a tedy ani složitosti organismu. Příčinou tohoto vysokého rozdílu mezi velikostmi genomů je množství tzv. nekódujících sekvencí DNA (Obr. 1), tedy sekvencí, které nejsou transkribovány do mRNA a následně nedochází k jejich translaci na konkrétní protein. Mezi nekódující sekvence patří introny, pseudogeny, regulační sekvence a repetitivní sekvence (Lodish *et al.*, 2016).

Introny jsou části RNA transkriptu, které musí být před translací odstraněny postranskričpními úpravami RNA. Ačkoli nekódují žádný protein, mohou po vystřihnutí sloužit jako funkční RNA molekuly, případně jako prostředky k regulaci genové exprese. Pseudogeny jsou nefunkční oblasti DNA vzniklé např. mutací původního genu, nebo neúplnou duplikací. Mezi regulační sekvence patří např. promotéry, zesilovače a jiné oblasti nacházející se v okolí genu, jejichž funkcí je regulace exprese daného genu (Lodish *et al.*, 2016).

2.2 Repetitivní DNA

Sekvenční motivy nacházející se v genomu ve stovkách až tisících kopiích tvoří velký podíl na celkové velikosti jaderného genomu u většiny eukaryotních organismů (Biscotti

et al., 2015). Ačkoli byly dříve považovány pouze za „genomické parazity“, dnes je vědecky dokázáno, že hrají důležitou roli ve vývoji genomů, genetické diverzitě a genové regulaci. Repetitivní DNA zahrnuje jak sekvence rozptýlené napříč genomem tzv. transpozibilní elementy, tak tandemové repetice tzv. satelitní sekvence DNA (viz Obr. 1). Oba typy sekvencí výrazně přispívají k odlišnosti ve velikosti genomů mezi rozdílnými druhy, přičemž v některých organismech tvoří více než 50 % celkového množství DNA (López-Flores a Garrido-Ramos, 2012).

Pojem satelitní sekvence DNA, zkráceně satDNA, je historicky spjatý s analyzováním DNA hustotně gradientní centrifugací, při které tandemově uspořádané sekvence tvořily satelitní bandy oddělené od zbytku genomické DNA (Kit, 1961). Typicky je v genomu organizována do dlouhých řad „head-to-tail“ spojených repetic (Charlesworth *et al*., 1994). Na základě délky monomerů (repetice, jejímž opakováním vznikne celá satelitní sekvence) a počtu jejich opakování lze satelitní DNA rozdělit na – mikrosatelity s délkou monomeru 2-5 bp s řádově 10–100 opakování, minisatelity s délkou monomeru 5–100 bp a délkou 0,5–30 kb, a satelitní DNA s rozličnou délkou monomerů (u většiny zvířat a rostlin 150-400 bp) a počtem opakování. Existují také případy, kdy jsou repetice uspořádány do vyšší struktury (z angl. higher-order repeat), ve kterých jednotlivé monomery tvoří sekvenční bloky, jejichž opakováním je daná satelitní sekvence tvořena (Willard, 1985). Na základě teoretických modelů evoluce satDNA, podloženými experimentálními daty, bylo ukázáno, že každá náhodná sekvence může být základem tandemové repetice (Garrido-Ramos, 2015).

V kontextu rozložení jaderné DNA se satDNA až na vyjímky vyskytuje ve shlucích v heterochromatinu (Plohl *et al*., 2012), který je lokalizován v centromerických a subtelomerických oblastech chromozómu (Adega *et al*., 2007). V rostlinné i živočišné centromerické DNA jsou zastoupeny mezidruhově výrazně homologní tandemové repetice (Melters *et al*., 2013). Oproti tomu subtelomerické repetice bývají často rodově specifické (např. TrsA u rodu *Oryza*, Cheng *et al*., 2001), či chromozómově specifické (např. WE 35 na chromozomu 5B u *Triticum Aestivum*, Ueng *et al*., 2000).

V genomu organismu se klasicky nachází více rodin satDNA. Jednotlivé rodiny se mohou lišit ve složení monomeru, velikosti, abundanci, rozmístění, nebo lokaci (Garrido-Ramos, 2017) a mohou být jak specifické pro daný organismus, tak konzervované v celé čeledi organismů, nebo řádu (Quesada del Bosque et al., 2013; Quesada del Bosque et al.,

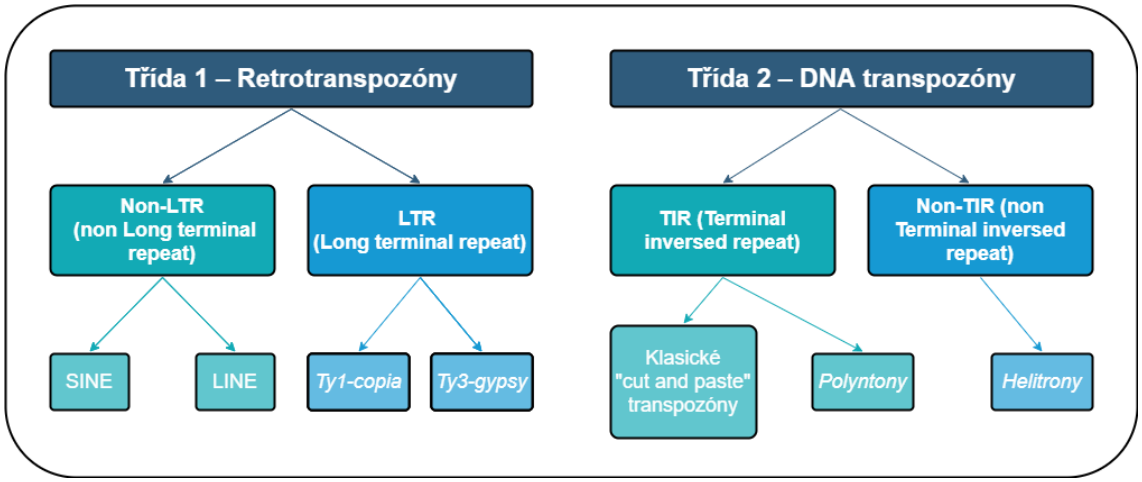
2014). Specifita rodin satDNA souvisí s rychlostí genomických změn v daných oblastech genomu. Změny mohou být provedeny specifickými mechanismy zodpovědnými za amplifikaci, či deleci repetic (Walsh, 1987). Mezi ně patří nerovnoměrný crossing over, replikační prokluzování (z angl. replication slippage) a amplifikace využitím otáčivé kružnice (z angl. rolling circle amplification).

Nicméně i rozdílné rodiny repetic mohou mít společné určité sekvenční motivy, jako dinukleotidy AA/TT, pentanukleotidy CAAAA, apod. (Mehrotra a Goyal, 2014). Sdílené motivy mohou být způsobeny podobnými mechanismy vývoje jednotlivých rodin repetitivních sekvencí, ale také mohou souviset s významností motivů při molekulárních mechanismech v genomu a na jejich základě lze určit např. jaké vlastnosti má chromatin, ve kterém jsou lokalizovány. Pro rozličné rodiny je typická přítomnost krátkých, přímých a obrácených repetic a krátkých palindromů. (Mehrotra a Goyal, 2014).

Funkce satDNA je spojována s různými buněčnými procesy, do nichž se zapojují oblasti genomu bohaté na satDNA – pohyb a párování chromozomů, chromozomové rekombinace, vazba dělícího vřeténka, přestavby chromozomů, vazba histonů aj. Všechny tyto děje určitým způsobem souvisí s evolucí a diferenciací karyotypu (Mehrotra *et al.,* 2013). Rovněž lze na základě přítomnosti pro daný organismus specifické satDNA rychle a spolehlivě daný organismus identifikovat. Repetitivní sekvence také mohou být vazebným místem pro specifické jaderné proteiny, s čímž souvisí tzv. histonový kód (konkrétní varianta posttranslačních modifikací histonových proteinů v jádře) a jeho regulační vliv na transkripci genů (Vogt, 1990).

Druhým typem repetitivní DNA jsou transpozibilní elementy, zkráceně TE, vyznačující se schopností měnit svou pozici napříč genomem pomocí tzv. transpozice, čímž mohou indukovat rozličné chromozomální mutace a alelickou různorodost (Wessler *et al*., 1995). Dle typu meziproduktu při transpozici se TE rozdělují (Obr. 2) na dvě skupiny – TE 1. třídy (mezičlánkem transpozice RNA) a TE 2. třídy (mezičlánkem transpozice DNA, Finnegan, 1989).

TE třídy 1, také nazývané retrotranspozóny, se v genomu pohybují pomocí „copy-and-paste“ procesu tzv. retrotranspozicí, při které je DNA elementu transkribována do mRNA a poté převedena zpátky do DNA na odlišném místě v genomu. Mechanismus přepisu mRNA do DNA je katalyzován reverzní transkriptasou, jež je každým retrotraspozónem



Obr. 2 Klasifikace transpozibilních sekvencí DNA.

kódována. Na základě odlišné struktury a integračního mechanismu lze TE třídy 1 rozdělit na dvě podtřídy – LTR retrotranspozóny a non-LTR retrotranspozóny (Obr. 2). Skutečnost, že obě rodiny sdílejí sekvenční motivy v kódující oblasti pro reverzní transkriptasu, naznačuje jejich společný evoluční původ.

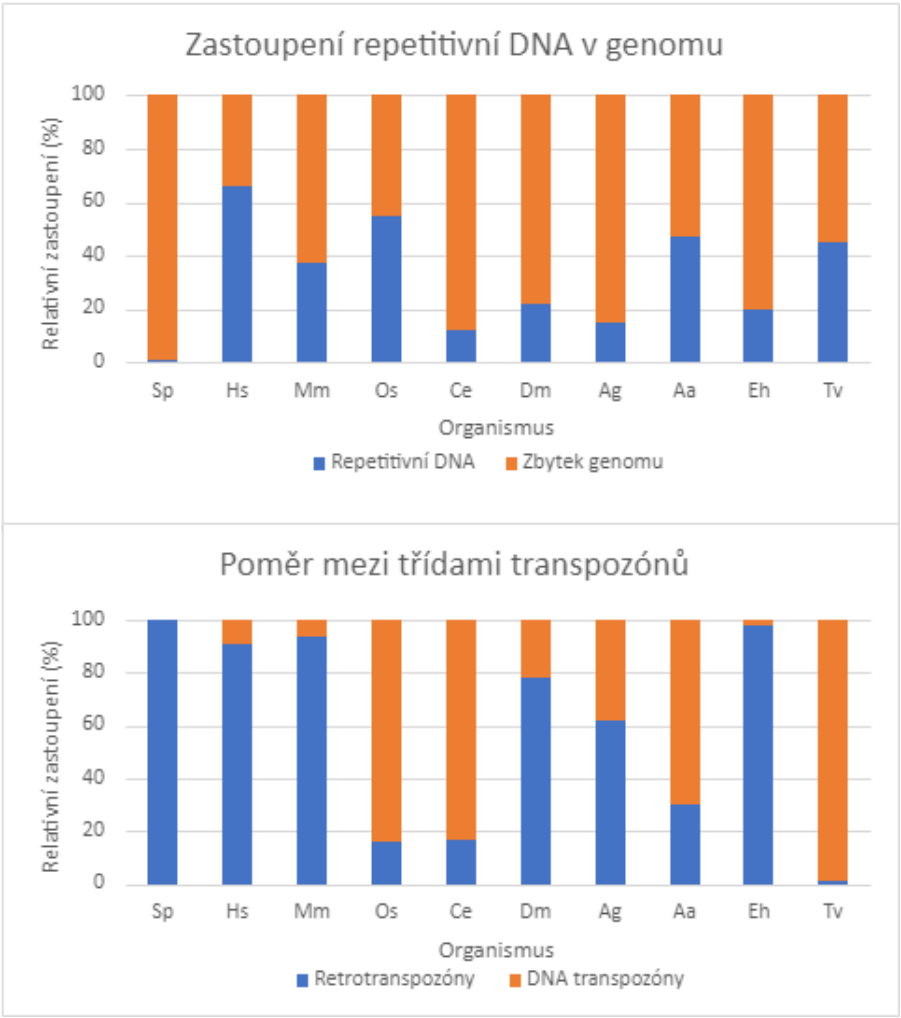
LTR retrotranspozóny mohou dosahovat délky od několika stovek po tisíce bp (Lee a Kim, 2014) a mají dlouhé terminální repetice na obou koncích, které mohou dosahovat délky několika stovek bp (z toho název LTR z angl. long terminal repeat). Mechanismus replikace je totožný s mechanismem množení retrovirů. LTR retrotranspozóny obsahují dva otevřené čtecí rámce – *GAG* a *POL*. *GAG* kóduje protein pro replikaci DNA a *POL* kóduje proteasu, integrasu, reverzní transkriptasu a RNasu H. Reverzní transkripce RNA intermediátu bývá zpravidla započata primerem na 3‘ konci tRNA, která se na RNA meziprodukt hybridizuje. Druhé vlákno DNA vzniká činností polymerasy, která rozpoznává dvě terminální repetice, což umožňuje její přechod mezi oběma konci RNA. Začlenění DNA do chromozomu je poté katalyzováno integrasou. LTR retrotranspozóny lze dále rozdělit na dvě podskupiny – *Ty1-copia* (*Pseudovirideae*) a *Ty3-gypsy* (*Metavirideae*). Rodiny se mj. odlišují pořadím kódujících oblastí (Wicker *et al*., 2007).

Non-LTR retrotranspozóny se kromě absence terminálních repetic odlišují také způsobem integrace. V tomto případě je cílové místo v DNA naštěpeno endonucleasou a 3‘ konec vzniklý rozštěpením je využit jako primer reverzní transkripce RNA meziproduktu přímo na řetězci DNA. Tento proces se nazývá cílově započatá reverzní transkripce (z angl. target-primed reverse transcription) a umožňuje nejen reverzní transkripci non-LTR retrotranspozónů, ale rovněž jiných RNA produktů (Dewannieux *et al*., 2003). Hlavními podtřídami non-LTR retrotranspozónů jsou SINE (z angl. short

interspersed nuclear elements) a LINE (z angl. long interspersed nuclear elements). Velikost SINE elementů je zpravidla 80–500 bp a jsou hojně zastoupeny v živočišných, převážně savčích genomech a nepříliš v rostlinných. Mezi známý příklad SINE patří *Alu* element, který se nachází v lidském genomu v 500 tisíci kopiích (Rowold a Herrera, 2000). Oproti tomu LINE elementy jsou delší, mohou dosahovat délky několika kb a také jsou převážně zastoupeny v živočišných genomech. Výjimkou je element *Del-2*, který je hojně zastoupen v rostlinách rodu *Lilia*.

Transpozibilní elementy 2. třídy, také zvané DNA transpozóny, pro transpozici nepotřebují RNA intermediát. Na základě odlišností ve struktuře a mechanismu transpozice je lze rozdělit (Obr. 2) na klasické transpozóny s „cut and paste“ transpozicí, *Helitrony* a *Polintony* (Wicker *et al*., 2007). Klasické transpozóny se vyznačují jednoduchou strukturou skládající se zpravidla z jediného čtecího rámce kódujícího transposasu s endonucleasovou a ligasovou aktivitou, umožňující vystřihnutí elementu a začlenění na jiné místo v genomu. Na obou koncích obsahují invertované repetice z toho zkratka TIR (z angl. terminal inverted repeats, Nassif *et al*., 1994). Vyznačují se specifickou sekundární strukturou, která umožňuje vazbu transposasy. *Helitrony* patří mezi vysoce rozšířené typy transpozónů a tvoří podstatnou část genomů některých rostlin (Du *et al*., 2006). Jejich mobilita v genomu je zprostředkována skrze mechanismus otáčivé kružnice. Obsahují sekvence kódující protein s centrální doménou homologní s proteiny umožňující replikaci otáčivou kružnicí u bakteriálních plazmidů, nebo fágů a C-terminální doménu podobnou PIF1 skupině u DNA helicas (Kapitonov a Jurka, 2006). Tento způsob replikace může vést ke tvorbě vysokého množství kopií a způsobovat rozdíly ve velikostech genomů, příkladem může být genom kukuřice (Lai *et al.*, 2005). *Polintony* se od ostatních DNA transpozónů odlišují velkým množstvím čtecích rámců (9–20), vyšší délkou (9–22 kb) a také schopností se replikovat. Obsahují totiž geny kódující jak DNA polymerasu (protein-primed DNA polymerase), tak integrasu podobnou retrovirové integrase (Krupovic a Koonin, 2016) a také TIR.

Rozmístění transpozibilních elementů v genomu není náhodné, naopak prokazuje rozličné stupně preferencí pro inzerci do různých částí genomu na základě funkce této části. U mnoha typů elementů také došlo k vyvinutí mechanismu k zaměření specifického



Obr. 3 Graf zachycující relativní množství retrotranspozonů a DNA transpozonů v rozdílných eukaryotických genomech (horní obrázek) a jejich poměr v daných genomech (dolní obrázek). Hodnoty představují procentuální zastoupení DNA transpozonů a retrotranspozonů vztažené k celkovému počtu transposibilních elementů v jednotlivých organismech. Vysvětlení zkratek organismů – Sc: *Saccharomyces cerevisiae*; Sp: *Schizosaccharomyces pombe*; Hs: *Homo sapiens*; Mm: *Mus musculus*; Os: *Oryza sativa*; Ce: *Caenorhabditis elegans*; Dm: *Drosophila melanogaster*; Ag: *Anopheles gambiae*, Aa: *Aedes aegypti*; Eh: *Entamoeba histolytica*; Ei: *Entamoeba invadens*; Tv: *Trichomonas vaginalis*. Vytvořeno na základě dat dle Maxvell, 2020; Guio a Gonzáles, 2019; Zuccolo *et al*., 2007; Sijen a Plasterk, 2003; Melo a Wallau, 2020; Arensburger *et al*., 2011; Lorenzi *et al*., 2010; Woehle *et al*., 2014.

lokusu, ve kterém se mohou namnožit, aniž by uškodili hostiteli (Sultana et al., 2017), např. několik retrotranspozónů se v odlišných organismech vyvíjelo nezávisle, avšak ve všech případech se nacházejí před 5‘ koncem úseků, které kóduje geny, tzn. že dochází k transkripci elementů, avšak neovlivňují transkripci genů (Spaller et al., 2016).

Výzkum v oblasti transpozibilních elementů prokázal jejich vliv na rozšiřování genomů, kvůli jejich schopnosti integrace do rozličných pozic v DNA, což z nich dělá zdroje nových prvků v genomu, které velkou měrou přispívají k vývoji organismů (Feschotte a Pritham, 2007). Velké množství příbuzných TE kopií rozprostřených po

genomu mohou vést k ektopické rekombinaci a duplikacím velkých DNA úseků. Nicméně většina efektů je spíše neutrální, nebo vede ke smrti jedince a pouze malá část TE nabízí organismu benefity (Bennetzen a Wang, 2014). Podíl obou typů transposibilních elementů u různých druhů organismů je znázorněn na Obr. 3.

2.3 Analýza repetitivních sekvencí

Existence repetitivní DNA v genomu eukaryot byla poprvé odhalena u myši v roce 1961 (Kit *et al.,* 1961; Seuoka *et al*., 1961). K experimentu byla využita hustotní gradientní ultracentrifugace DNA s užitím chloridu cesného, při níž byly identifikovány dva DNA bandy, z nichž jeden podléhal renaturaci rychleji než zbytek jaderné DNA myši. U jednoduchých genomů obsahujících pouze unikátní DNA sekvence k tomuto jevu však nedocházelo, což vedlo k důkazu přítomnosti repetitivních úseků (Waring a Britten, 1966). Následně byla vynalezena Cot analýza, založená na principech DNA renaturace, kdy DNA sekvence reasociuje rychlostí, která je přímo úměrná počtu výskytů této sekvence v genomu, čímž bylo dokázáno, že rozdílná část genomu každého eukaryotního organismu je tvořena repetitivními elementy lišícími se počtem repetic (Britten *et al*., 1974).

Od té doby bylo k analýze repetitivních sekvencí využita široká škála metod založených na rozdílných principech. Mimo jiné se jednalo o aplikaci restrikčních endonukleas na genomickou DNA, následná separace elektroforézou a analýza výsledných fragmentů (Singer, 1982). K přibližné kvantifikaci a organizaci v genomu byly využívány techniky dot-blot a Southern blot. Metody jsou založeny na imobilizaci fragmentů DNA na membráně a hybridizaci značených sond na fragmenty odpovídající satDNA. Na základě porovnání množství značených fragmentů vůči standardu lze zjistit jejich relativní množství (Garrido-Ramos *et al*., 1995). Velký průlom v lokalizaci satDNA v genomu umožnily techniky *in situ* hybridizace (Gall, 2016), především pozdější metoda FISH (z angl. Fluorescent *in situ* Hybridization). Využitím těchto postupů bylo možno pomocí značených sond identifikovat satDNA přímo na metafázních chromozomech (Pardue a Gall, 1970).

Důležitým milníkem v analýze struktury repetitivních sekvencí byl vývoj metod nové generace sekvencování (z angl. next-generation sequencing, zkr. NGS), avšak bylo nutné vymyslet sofistikovanější metody pro zpracování sekvenačních dat. (Garrido-Ramos, 2015). V případě sekvencování metodou Illumina dochází k tvorbě přesných, ale

krátkých sekvenačních readů, u nichž nastává problém při skládání (z angl. assembly) repetitivních oblastí. Repetitivní charakter sekvencí znemožňuje jednoznačné složení sekvenačních readů v případě, kdy je repetice výrazně delší než read, protože nelze dopředu určit, kolik opakování daná repetice obsahuje. Navíc PCR kroky, které jsou součástí Illumina sekvencování, mohou zvýhodňovat sekvence s velkým množstvím GC nukleotidů, což může vést k nepřesným informacím o velikosti repetitivních úseků. Tento problém lze překlenout dvěma způsoby – využitím algoritmů zpracovávajících hrubá neposkládaná sekvenační data nebo využitím nových metod sekvencování, které tvoří delší sekvenační ready a jejichž skládání je jednoznačnější a méně výpočetně náročné (Lower *et al*., 2018).

K analýze repetitivní DNA z hrubých sekvenačních dat vzniklo několik algoritmů, lišících se strategií tří hlavních kroků analýzy – identifikace readů příslušejících repetitivním sekvencím, přiřazení těchto readů do jednotlivých skupin odpovídajících dané repetici a kvantifikace zastoupení daných repetic v genomu a jejich variací (Lower *et al*., 2018).

V případě, že není potřeba vyhledávat nové repetitivní DNA, ale pouze repetice již objevené u jiného organismu/vzorku, lze využít porovnání se sekvencemi známých repetic, tento způsob je využit např. v softwaru Alpha-CENTAURI (Sevim *et al*., 2016). Pokud je cílem analýzy *de novo* identifikace repetic, existují dva základní principy – vzájemné porovnání (z angl. alignment) readů s nízkým pokrytím (lze dosáhnout např. náhodným zredukováním počtu readů vzniklých sekvencováním), nebo vyhledávání rekurentních motivů mezi ready. Tento přístup funguje, protože při nízkém pokrytí se statisticky vyskytují ve více kopiích pouze repetitivní sekvence. Na tomto principu jsou založeny algoritmy RepeatExplorer (Novák *et al*., 2013) a TAREAN (Novák *et al*., 2017). Nevýhodou tohoto přístupu je nízká citlivost pro málo zastoupené repetice. Oproti tomu hledáním rekurentních motivu, např. pomocí rozdělení jednotlivých readů na k-tice lze správně identifikovat i málo zastoupené sekvence, ale délka detekovatelného monomeru je limitována délkou readů. Na tomto přístupu je založen např. algoritmus k-Seek (Wei *et al*., 2014).

Pro následné správné zařazení readů do repetic je často využívána reprezentace pomocí grafů vytvořených na základě sekvenční podobnosti readů. Každá repetice poté přísluší jednomu grafu a jeho další analýzou lze získat i statistické údaje o daném typu

repetice. Implementace algoritmů RepeatExplorer i TAREAN tuto reprezentaci zahrnují. Existují také algoritmy, které kombinují různé přístupy k základním krokům analýzy repetitivních sekvencí a nabízejí pokročilé informace, mezi ně patří např. satMiner (Ruiz-Ruano *et al*., 2016).

S vývojem technologií přichází i nové sekvenační metody, jejichž délka readů může dosahovat až několik desítek, či stovek kb, tedy řádově stokrát více, než je tomu v případě Illumina sekvencování. Mezi tyto nové technologie se řadí Pacific Bioscience sekvencování (Khost *et al*., 2017) a Oxford Nanopore sekvencování (Jain *et al*., 2018). S takto dlouhými ready lze překlenout dlouhé repetitivní úseky a využitím vylepšených algoritmů poté složit ready do správné podoby (Chakraborty *et al*., 2016).

Další možností, jak vylepšit výsledky sestavení readů, je využití optického mapování (Lam *et al*., 2012). Technika je založena na detekci fluorescenčně značených úseků na celém nataženém vlákně DNA, tímto způsobem lze vytvořit fyzickou mapu a získat informace o celkové struktuře genomu. Nicméně je tato technika méně citlivá na krátké repetiční monomery, protože v případě, že se vyskytují vedle sebe dochází ke splývání fluorescenčního značení. Při výzkumu exprese satDNA, nebo její asociace s chromatinem lze aplikovat metody RNA sekvencování (RNA-seq) nebo chromatinové imunoprecipitační sekvencování (ChIP-seq, Barski *et al*., 2007).

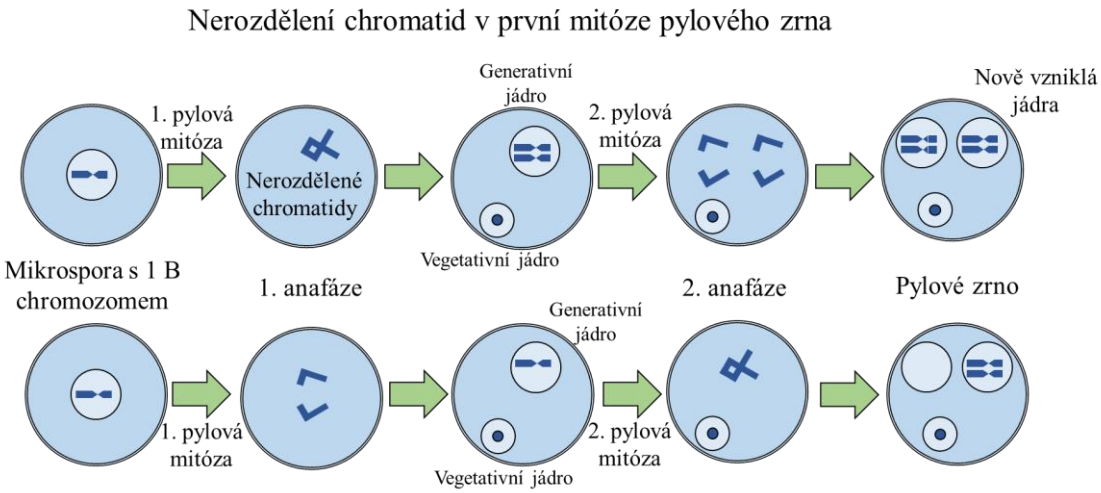
2.4 B chromozomy

B chromozomy jsou postradatelnou součástí karyotypu vykazující nemendelistický způsob přenosu do dalších generací. Jejich počet se zpravidla liší i u zástupců jedné populace a může se v závislosti na organismu pohybovat v rozmezí jednotek až desítek (Camacho *et al*., 2000). Jejich dalším rysem je absence schopnosti rekombinace se standardními A chromozomy (Longley, 1927). Předpokládá se, že jsou B chromozomy obsaženy v 15 % eukaryotických organismů (Beukeboom, 1994), z nichž většinu tvoří rostliny (Jones *et al*., 2008). Typicky nemají vliv na funkci organismu, ale jejich zvyšující se počet může vést k destabilizaci genomu a mít zhoubné následky na organismus (Houben, 2017).

Sekvence B chromozomů mohou být originální danému B chromozomu, ale častěji pochází buď z ostatních B chromozomů nacházejících se v genomu organismu, nebo z klasických chromozomů (A chromozomů), avšak ancestrální sekvence může pocházet

také z A chromozomu jiného druhu organismu (Perfectti a Werren, 2001). Přenos sekvencí na nový B chromozom může být zapříčiněn aktivitou transpozibilních elementů, interchromozomální meiotickou rekombinací, či nedokonalou DNA opravou (Banaei-Moghaddam *et al*., 2013). Sekvence mohou pocházet jak z protein kódujících, tak nekódujících genů. Valná většina protein kódujících genů B chromozomů pochází z A chromozomů, nicméně je velké množství z nich určitým způsobem degenerováno, důsledkem neúplné kopie ancestrálního genu, chybějících exonů, nebo jiné esenciální oblasti, či obecně nízkou homologií s původním genem (Ruban *et al*., 2017). Většina sekvencí na B chromozomech je nekódujících a pocházejí hlavně z transpozibilních elementů a satelitních repetic (Camacho *et al*., 2000). Skutečnost, že se tyto typy sekvencí nacházejí ve velkém počtu v heterochromatinu centromerických oblastí chromozomů, napovídá tomu, že původ B chromozomu souvisí s duplikací A chromozomu, který ztratí euchromatinové oblasti a zůstane z nich pouze centromerická a pericentromerická oblast (Obr. 5). Následný růst B chromozomů již může být způsoben akumulací sekvencí vlivem aktivity transpozibilních elementů, tyto sekvence poté podléhají mutacím a jiným chromozomovým přestavbám a tím dochází k sekvenční různorodosti B chromozomů (Marques *et al*., 2018).

Jak již bylo naznačeno, u B chromozomů nedochází k rovnoměrné segregaci. Proces převodu chromozomů na dceřiné buňky, jehož poměr je vyšší než 0,5 (v případě klasické Mendelovské dědičnosti je hodnota 0,5) se nazývá drajv. Drajv může probíhat jak v pre-meotické, meiotické, tak post-meiotickém dělení v závislosti na typu organismu (viz Obr.



Obr. 4 Schéma zobrazující možnosti nerovnoměrného rozdělení B chromozomu do dceřiných buněk pylového zrna. Vytvořeno dle Jones *et al*., 2008.

4). Rovněž počet B chromozomů, které jsou organismem tolerovány, se liší, u kukuřice se může v jádře nacházet až 34 B chromozomů. V případě žita (*Secale cereale L.*) je drajv způsoben rozšířením pericentromerické oblasti B chromozomů o NCR (z anglického non-disjunction control region), který zabraňuje rozdělení sesterských chromatid B chromozomů v první mitóze pylových zrn (Endo *et al.,* 2008, viz Obr. 4). Tím dochází k tzv. asymetrickému rozdělení chromozomů (Twell, 2010). Avšak v případě, že B chromozomy NCR postrádají, proběhne mitóza řádně a počet B chromozomů je u obou sesterských buněk stejný. Oproti tomu v kukuřici dochází k drajvu kvůli nerozdělení B chromatid při druhém mitotickém dělení pylového zrna (Obr. 4), v důsledku čehož dochází k zanechání dvou kopií B chromozomů v jedné ze dvou spermií (Houben 2017). Spermie obsahující B chromozomy také přednostně oplodní vajíčko (Lamb *et al.,* 2007).

B chromozomy mohou mít výrazní vliv na organismus i v případě, kdy neobsahují geny kódující proteiny, např. u *Nasonia vitripennis* jsou B chromozomy (nazývané PSR chromozomy z anglického paternal sex ratio) přeneseny pouze skrze spermii a způsobují odstranění jedné sady A chromozomů v průběhu první mitózy zygoty (Nur *et al*., 1988). Výsledkem je přeměna samičí zygoty na samčí embryo (Aldrich a Ferree, 2017).

2.5 Repetitivní sekvence B chromozomů u rostlin

Typickým příkladem repetitivních sekvencí B chromozomů jsou 5S a 45S rDNA satelitní repetice, pro které je typický vysoký stupeň mezidruhové evoluční stability a také se v genomu nacházejí ve velkém počtu opakování. Tyto satDNA byly objeveny ve velkém počtu rostlin i zvířat, např. *Brachycome dichromosomatica* (Donald *et al*., 1997), *Crepis capillaris* (Maluszynska a Schweizer, 1989), nebo *E. plorans* (López-León *et al*., 1994). V některých případech dokonce dochází k transkripci těchto sekvencí (Leach *et al*., 2005).

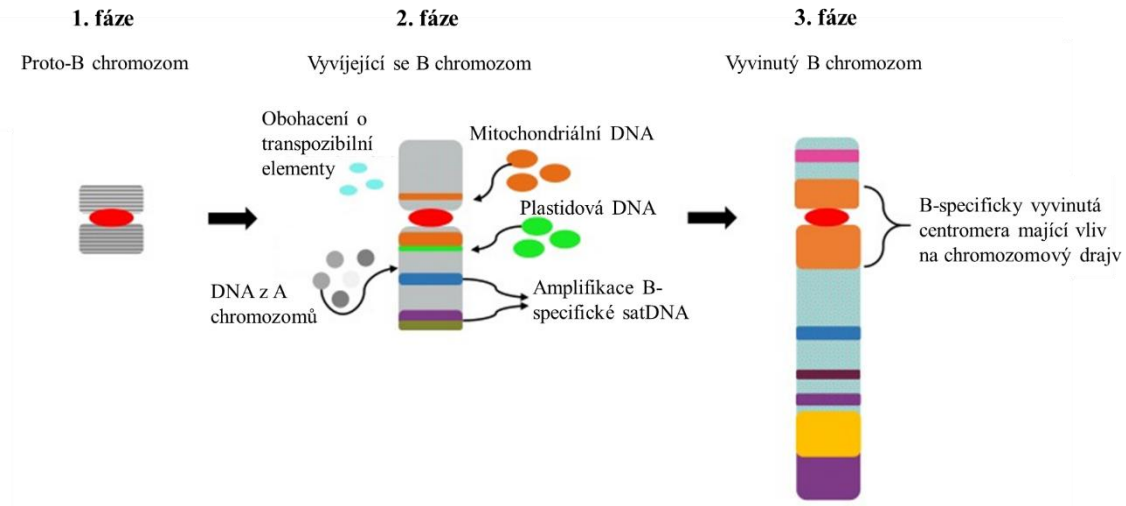
První B-specifické satelitní repetice u rostlin byly objeveny v genomu žita (*Secale cereale*). Jedná se o sekvence označené E3900 a D1100, které se vyvinuly *de novo* (Blunden *et al*., 1993) s atypickými délkami monomerů (1,1 kb a 3,9 kb). Sekvence vznikly složením fragmentů různých DNA elementů pravděpodobně vlivem chromozomových přestaveb (Langdon *et al*., 2000). Sekvencováním a *in silico* analýzou genomu žita byly objeveny i další B-specifické repetice, především satDNA (Klemme *et*

al., 2013) a také byly identifikovány oblasti z A chromozomů, ze kterých je B chromozom vytvořen – přestavba chromozomů 3RS a 7R, fragmenty genů a repetic z ostatních A chromozomů a inzerce organelové DNA (Martis *et al*., 2012). Repetice jsou především akumulovány v rozšířené oblasti pericentromery a na konci dlouhého ramene chromozomu (Klemme *et al*, 2013), kde se nachází NCR.

Obdobně v genomu kukuřice (*Zea mays*) se nacházejí B-specifické repetitivní sekvence (Stark *et al*., 1996). Jednou z nich je repetice s označením ZmBs, mající přibližnou délku monomeru 1,4 kb. Nachází se okolo i uvnitř centromery a rovněž na konci dlouhého ramene chromozomu (Alfenito a Birchler, 1993). Další B-specifickou repeticí je CL-1 s přibližnou délkou monomeru 1,5 kb, ta se nachází ve třech heterochromatických blocích dlouhého ramene chromozomu (Cheng a Lin, 2003). Ani jedna z těchto repeticí není homologní k sekvencím na A chromozomech, avšak zde byla detekována transpozice retrotranspozónu a také MITE elementu, které patrně stáli u zrodu těchto repetic (Cheng a Lin, 2004). Specifické repetitivní sekvence jsou tedy akumulovány v podobných oblastech jako v případě žita.

V případě všelichy (*Brachycome dichromosomatica*) byly v genomu identifikovány dva typy B chromozomů – větší, které jsou somaticky stabilní a menší, které jsou naopak somaticky nestabilní. Větší B chromozom obsahuje specifickou tandemovou repetici Bd49, jež je homologní s repeticemi přítomných na A chromozomech příbuzných druhů, kdežto v A chromozomech *B. dichromosomatica* je této repetici podobných jen několik oblastí (Leach *et al*., 1995). Menší B chromozom obsahuje specifické repetice Bdm29 a Bdm54, přičemž repetitivní sekvence podobné Bdm29 se nacházejí i ve velkém B chromozomu a v ostatních druzích rodu *Brachycome*, což ukazuje na velkou konzervovanost a rozptýlenost těchto repetic napříč genomy (Houben *et al*., 1997). Také bylo ukázáno, že genomická organizace malého B chromozomu neodpovídá žádné oblasti z A chromozomu, což znamená, že chromozom nemohl vzniknout přímou duplikací (Houben *et al*., 2001).

Zatím bylo provedeno malé množství studií týkajících se akumulace transpozibilních elementů na B chromozomech rostlin, nicméně se předpokládá, že jsou B chromozomy ideálním místem pro jejich transpozici a následné rozšíření. U žita je několik typů transpozónů rozšířeno v mnohem větším, nebo naopak menším poměru, než tomu je u A chromozomů. Příkladem může být element Sabrina, který ačkoli je abundantní ve všech



Obr. 5 Schéma vývoje B chromozomu spolu se zdroji akumulujících repetitivních sekvencí. Převzato a upraveno podle Marques *et al*., 2018.

Triticae a transkripčně neaktivní v žitě (Shirasu *et al*., 2000), tak je v A chromozomech zastoupen v mnohem větším množství než v B chromozomech. Naopak kopií elementu Revolver je v B chromozomech žita mnohem více, v porovnání s jeho přítomností v A chromozomech (Kleeme *et al*., 2013). Příčinou B-specifické akumulace elementů může být nízká míra meiotického crossing-overu, který nastává, kvůli menšímu počtu chiasmat v bivalentních B chromozomech (Jiménez *et al*., 2000). Následkem crossing-overu totiž dochází k odstranění mobilních elementů (Charlesworth *et al*., 1994).

Omezení meiotických rekombinací může obecně vést k akumulaci repetitivních sekvencí a také k počátku nezávislého vývoje B chromozomů (Jiménez *et al*., 1997). Přítomnost rychle se rozšiřujících repetitivních sekvencí podporuje náchylnost k dalším strukturním modifikacím potřebným k zavedení a amplifikaci nových B-specifických repetic (Marques *et al*., 2018). Druhou možnou příčinou akumulace repetitivních sekvencí je redukovaný selektivní tlak na B chromozomy v souvislosti s jejich neesenciální povahou. Tyto vlastnosti představují ideální prostředí pro nekódující, rychle se vyvíjející sekvence jako satDNA, organelová DNA, nebo transpozibilní elementy, pro které je příznačná přítomnost v regionech nepodléhající rekombinaci (Marques *et al*., 2018).

Obecně lze model akumulace a vývoje repetitivních sekvencí na rostlinných B chromozomech charakterizovat ve třech krocích (Obr. 5). Prvním krokem je odvození proto-B chromozomu od původního A chromozomu způsobené různými translokacemi a duplikacemi. B chromozom v této fázi sdílí sekvence s A chromozomem a nutně musí

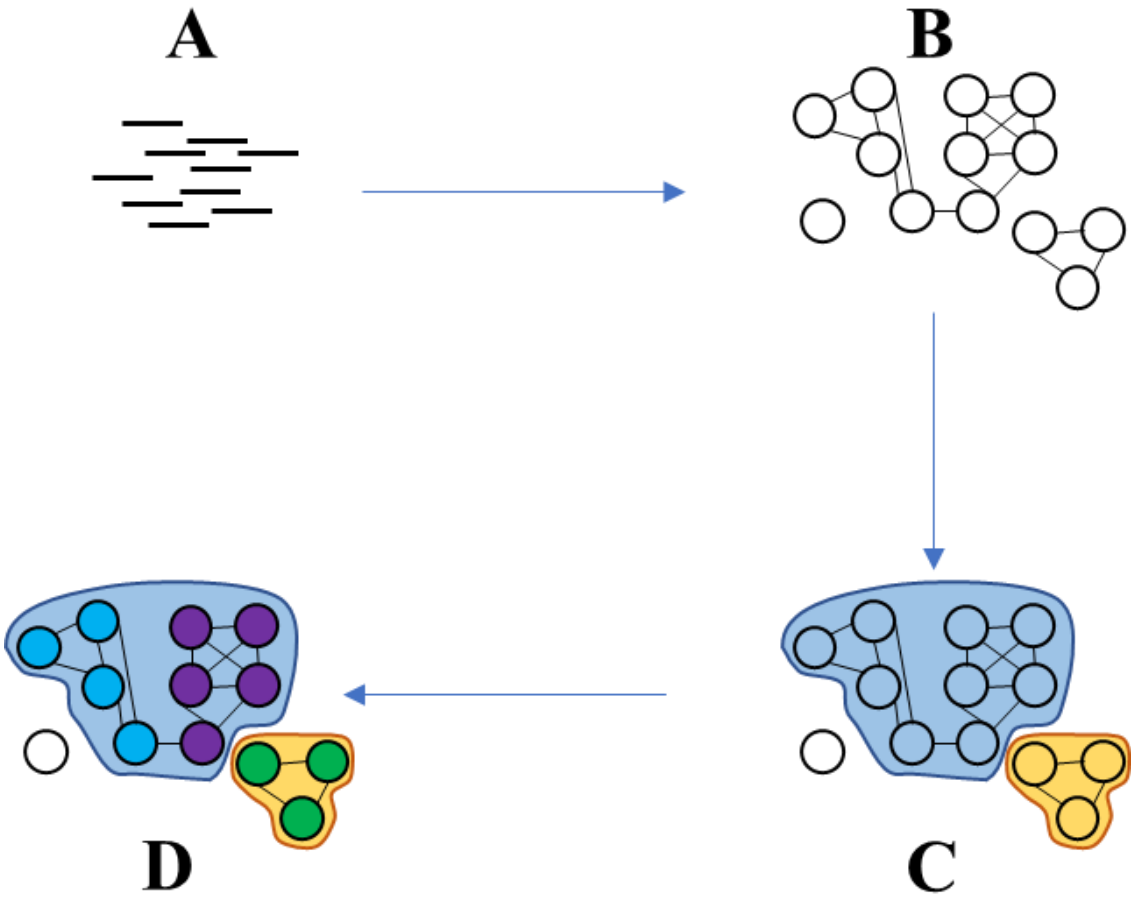
obsahovat funkční centromerní oblast pro zaručení mitotického a meiotického dělení. Druhým krokem je postupné umlčování genů a akumulace repetitivních sekvencí v důsledku nedokonalé rekombinace a nízkého selekčního tlaku. Dochází také k vývoji specifického drajvu a mechanismu dělení, důsledkem čehož zůstávají B chromozomy i v dalších generacích. V poslední fázi dochází k ústálení B-specifických repetic a sekvencí s únosným vlivem na organismus. Nicméně dokud nedojde k chromozomovému odstranění, může neustále akumulovat další sekvence, procházet mutacemi a vylepšovat mechanismus drajvu. Poznatky z genomů žita a kukuřice ukazují, že s drajvem souvisí vývoj centromerické a pericentromerické oblasti, který u těchto rostlin zapřičiňuje nerozdělení chromatid v prvním nebo druhém mitotickém (Obr. 4) dělení pylových zrn (Houben, 2017).

2.6 RepeatExplorer

Jak již bylo naznačeno, softwarový balík RepeatExplorer slouží k *de novo* identifikaci a podrobné analýze repetitivních sekvencí na základě zpracování sekvenačních readů s pokrytím menším jak 0,5. Je založen na klastrování readů na základě jejich vzájemné podobnosti a je implementován v programovacím jazyce R (Dessau a Pipper, 2008).

Úvodním krokem algoritmu je vzájemné párové porovnání readů a uložení všech párů, jejichž podobnost překročila určitý předem definovaný práh. Následně dojde na základě těchto informací k vytvoření grafu (Obr. 6B), v němž vrcholy reprezentují jednotlivé ready. Dva vrcholy jsou spojeny hranou, pokud byly ready příslušné těmto vrcholům klasifikovány jako dostatečně podobné. Hrana je vždy ohodnocena skórem podobnosti dvou vrcholů, které spojuje. Protože se pracuje s genomickými daty s nízkým pokrytím, případ, kdy jsou dva vrcholy spojené hranami odpovídá s největší pravděpodobností tomu, že pokrývají repetitivní sekvenci. Naopak vrcholy odpovídající readům pokrývajícím úseky nachazející se v genomu v jedné kopii zůstanou v grafu bez spojení hranami (Novák *et al*., 2010).

Vytvořený graf lze rozdělit do tzv. souvislých komponent – podmnožin vrcholů, které tvoří souvislý graf (mezi libovolnou dvojicí vrcholu existuje posloupnost vrcholů taková, že mezi dvěma po sobě jdoucími vrcholy je hrana) a které jsou maximální (nelze přidat žádný vrchol, aniž by se porušila souvislost, viz Obr. 6C). Identifikace souvislých komponent je hlavní ideou algoritmu softwaru tclust (Pertea *et al.*, 2003), jenž byl také využíván k analýze repetitivních sekvencí (Macas *et al*., 2007). Tento přístup v ideálním



Obr. 6 Schéma základních kroků algoritmu RepeatExplorer pro identifikaci repetitivních sekvencí. A – vstupní sekvenační ready. B – na základě informací z párového porovnání vstupních readů je vytvořen graf, jednotlivé vrcholy reprezentují ready, hrany spojují dva vrcholu, pokud jsou ready dostatečně podobné. C – dochází k identifikaci souvislých komponent v grafu, naznačeno podbarvením daných uzlů a hran. D – dochází k vyhledání komunit, které reprezentují daný typ repetitivní sekvence, zaznačeno obarvením vrcholů.

případě vede k identifikaci jednotlivých typů repetitivních elementů, které jsou totožné se souvislými komponenty. V reálných případech ale dochází k částečné homologii repetitivních sekvencí, která vede ke vzniku souvislých komponent obsahujících ready z odlišných typů repetic (Novák *et al*., 2010).

Z tohoto důvodu je v algoritmu implementována analýza grafu prostřednictvím hierarchického aglomeračního algoritmu (Clauset, et al., 2004), jehož funkcí je vyhledání komunit (Girvan a Newman 2002). Časová složitost algoritmu je $O(md\ log\ n)$, kde n je počet vrchol, m je počet hran a d je hloubka grafu reprezentujícího komunitu. Komunita je definována jako podmnožina vrcholů grafu, u níž je hustota propojení hranami mezi vrcholy uvnitř podmnožiny větší než se vrcholy mimo podmnožinu (Obr. 6D). K vyhledávání optimálního rozdělení grafu na komunity je využit greedy algoritmus, který identifikuje rozklad, jehož komunity budou mít maximální modularitu. Modularita je

kvalitativní mírou grafového klastrování, jejíž hodnota odkazuje na frekvenci hran mezi vrcholy dané komunity vztaženou k očekávané frekvenci náhodně pospojovaného grafu (Newman, 2006; Newman a Girvan, 2004). Hodnota modularity se blíží 0 v případě, kdy se frekvence hran nepříliš liší od náhodně sestaveného grafu a blíží se 1, když se naopak liší výrazně ve prospěch větší frekvence. Výsledné klastry reprezentující jednotlivé typy repetic jsou totožné s optimálním rozdělením grafu na komunity (Novák et al., 2010).

Každý klastr je následně analyzován za účelem získání podrobnějších informací o složení a početnosti dané repetice. Počet readů přiřazených danému klastru proporčně odpovídá zastoupení repetice v celém genomu, protože každý read odpovídá náhodné oblasti DNA a výskyt stejné sekvence ve více readech přímo souvisí s její genomovou četností. Složení repetice pak odpovídá konsenzuální sekvenci daného klastru. Další statistické údaje lze získat topologickou analýzou klastrů. Některým typům repetitivních sekvencí totiž odpovídají specifické hodnoty vlastnosti tvaru klastru. Lze vypočítat např. průměr grafu jako nejdelší cestu (posloupnost vrcholů takových, že mezi následujícími vrcholy existuje hrana) mezi libovolnýma dvěma vrcholy, nebo hustotu grafu jako poměr počtu hran vůči maximálnímu možnému počtu hran, anebo maximální stupeň jako maximální počet hran vedoucích do jednoho uzlu. Bylo dokázáno, že průměr grafu je přímo úměrný délce repetitivního elementu, a že vysoká hustota a maximální stupeň grafu jsou typické pro krátké tandemové repetice (Novák et al., 2010).

K prohlížení 3D struktury klastrů byl vyvinut balíček SeqGrapheR v programovacím jazyce R. Ten nabízí uživateli jednoduché grafické rozhraní s možnostmi interaktivní vizualizace klastrů. Jeho implementace zahrnuje využití programu GGObi z balíčku rggobi (Lawrence et al., 2009; Swayne et al., 2003). Pomocí balíčku SeqGrapheR lze také vybírat skupiny readů z grafu a prohlížet jak jejich pozici v grafu, tak výsledky skládání sekvencí a vyhledávání na základě podobnosti (Novák et al., 2010).

RepeatExplorer je implementován v prostředí Galaxy (webová platforma poskytující nástroje pro vědeckou analýzu, Afgan et al., 2018) a dostupný na webové adrese https://repeatexplorer-elixir.cerit-sc.cz/galaxy/. Prostředí Galaxy umožňuje základní úpravu sekvenačních dat (ořezání, filtrování na základě kvality apod.). Kromě identifikace repetitivních sekvencí nabízí RepeatExplorer nástroje k identifikaci, extrakci a analýze oblastí DNA patřících ke konzervovaným doménám retrotranspozónů a také metody k úpravě FASTQ a FASTA souborů (Novák et al., 2007).

2.7 TAREAN

Z tvaru klastrů lze rovněž odvodit, jestli reprezentuje tandemové repetice (kružnicový tvar klastru), nebo transpozibilní elementy (lineární tvar klastru). Za účelem pokročilejší analýzy repetitivních sekvencí a identifikace satelitní DNA je součástí komparativní analýzy také aplikace algoritmu TAREAN (Novák *et al.*, 2017). Hlavní ideou algoritmu je zpracováním klastrů (vzniklých totožným přístupem jako u algoritmu RepeatExplorer) a identifikace jejich cirkulárních úseků.

Cirkulární charakter klastru je detekován následovně. Z klastru je vytvořen nový graf Ω se zápornými a negativními hranami, jehož vrcholy reprezentují klastrová čtení, hrany odpovídají skóre podobnosti sekvencí dvou čtení a znaménko hrany je určeno orientací čtení (+ pro forward, - pro reverse). V takto vytvořeném grafu Ω je vyhledána nejmenší kostra Ω_{MKG} (podgraf obsahující všechny vrcholy, přičemž mezi libovolnými dvěma vrcholy existuje cesta, nejmenší v kontextu součtu ohodnocení všech hran podgrafu). Následuje průchod kostrou do hloubky, kde v případě, že vrcholu předchází hrana s negativní váhou, dojde k nahrazení sekvence čtení reprezentovaného vrcholem jeho reverzním komplementem. V takto vzniklém grafu G je vyhledána největší silně spojitá komponenta G_{NSK} (spojitá komponenta, ve které existuje cesta mezi libovolnými vrcholy). Na základě poměru počtu vrcholů G_{NSK} vůči G je vypočítán index spojité komponenty C. Graf má cirkulární charakter v případě, že se index rovná jedné.

Ačkoli index spojité komponenty slouží k odfiltrování netandemových repetic, nestačí k určení, zda je daná repetice satelitní, nebo tvoří mini, či mikro satelity. Z tohoto důvodu je zaveden další parametr, index párové úplnosti P, který je definován jako poměr úplných párů čtení (v klastru se nacházejí oba směry čtení) vůči počtu úplných a neúplných párů. Hodnota indexu blížící se jedné je typická pro tandemové satelitní repetice, protože z důvodu jejich délky (až několik Mb) se většina párů čtení nachází v dané repetici (klastru) v úplné podobě, na rozdíl od mini a mikro satelitů.

Sekvence monomerů repetic je určena pomocí k-mer analýzy, ve které dochází k paralelnímu rozdělení sekvencí všech čtení (díky předchozí úpravě jsou pouze ve forward směru) na všechny podsekvence délky k (tzv. k-mery), pro $k = 11–27$ bp. Jednotlivé k-mery jsou seřazeny podle zastoupení v daných čteních. Nejhojněji zastoupený k-mer je následně vybrán a je podle něj vytvořen de Bruijnův graf B, ve kterém se váhy vrcholů rovnají zastoupení k-meru ve čtení, který tento vrchol

reprezentuje. V grafu B je poté vyhledána největší silně spojitá komponenta B_{NSK}. V případě, že silně spojitá komponenta v grafu neexistuje, nebo součet vah vrcholů v ní nedosahuje předem definované hraniční hodnoty p_{km} (její optimální hodnota byla testována empiricky a nastavena na 0.225), tak je k současnému grafu přidán další nejvíce zastoupený k-mer a proces se opakuje, dokud nejsou podmínky splněny. Varianty sekvencí monomeru jsou poté určeny z kružnic v de Brujnově grafu pomocí seřazení (z angl. alignment) k-merů. Tímto způsobem je určena konsenzuální sekvence monomeru spolu s poziční pravděpodobnostní maticí.

Mimo to dokáže algoritmus TAREAN odlišovat 45S a 5S ribozomální RNA, pro které je rovněž typické tandemové uspořádání, pomocí vyhledání podobností se sekvencemi ve REXdb Viridiplantae rDNA databázi. Podobně dochází k porovnání konsenzuální sekvence monomerů vůči sekvenčním rysům LTR-retrotranspozonů (místo pro vazbu primeru, čtecí rámec pro kódování proteinů) k odlišení transpozonální DNA od tandemových repetic

3 EXPERIMENTÁLNÍ ČÁST

3.1 Rostlinný materiál

Semena *Sorghum purpureosericeum* (Hochst. E A.Rich) Schweinf. & Asch., u kterých nebyla dopředu známa přítomnost B chromozomů, byla dodána organizací ICRISAT (z angl. International Crop Research Institute for the Semi-Arid Tropics). Před výsevem byly ze semen odstraněny tvrdé slupky a následně byla semena přes noc namočena ve vodě. Naklíčení semen proběhlo v Petriho miskách v termálním inkubátoru s fotoperiodou nastavenou na 8 hodin světla při 29 °C a 16 hodin tmy při 25 °C. Takto připravená sadba byla zasazena do květináčů s 10cm průměrem, které obsahovaly zeminu smíchanou s pískem v poměru 2:1. Kultivace probíhala za stejných podmínek jako klíčení semen.

3.2 Izolace pylové DNA a její sekvencování

Haploidní buněčná jádra z pylových zrn rostlin, které obsahovaly B chromozomy (dále označovány jako B+) a s absencí B chromozomů (dále označovány jako B-) (identifikátor rostlinného materiálu – IS 18947) byla izolována pomocí průtokové cytometrie, při které bylo 9000 buněčných jader (přibližně 20 ng DNA) z každé rostliny separováno do 0,5ml zkumavek s 40 µl destilované vody. Z každého takto připraveného vzorku byla následujícím způsobem vytvořena sekvenační knihovna. Proteiny nacházející se v jádrech byly degradovány ošetřením vzorků 1,8 µl Proteinasy K (10 mg/ml). Po 18 hodinách při teplotě 50 °C byla aktivita Proteinasy K zastavena působením teploty 85 °C po dobu 30 minut. Následně byly vzorky zamrazeny na teplotu -80 °C. Fragmentace DNA proběhla v šesti třicetisekundových cyklech sonikátoru Bioruptor Plus (Diagenode, Denville, NJ, USA) při výkonu nastaveném na 300W. DNA byla poté purifikována využitím purifikační soupravy AmpureXP (Beckman Coulter, Brea, CA, USA). Jednotlivé sekvenační knihovny byly připraveny využitím soupravy NEBNext® Ultra™ II pro Illumina sekvencování (Ipswich, MA, USA) v 16 PCR cyklech. Z takto vzniklých knihoven byly vybrány ty, jejichž velikost se nacházela v rozmezí 700–1000 bp využitím nástroje Blue Pippin (Sage Science, Beverly, MA, USA). Sekvencování knihoven proběhlo v sekvenátoru NovaSeq 6000 za vzniku 2×250 bp dlouhých pair-end čtení.

3.3 Infrastruktura MetaCentrum

Při manipulaci a zpracování dat byl využit server IBM x3850 X5 zapojený do národní infrastruktury Metacentrum, jenž poskytuje paměťové a výpočetní zdroje pro práci s nainstalovaným softwarem. Mezi využitý software patřil *Trimmomatic, BWA* a *IGV*.

3.4 Úprava sekvenačních dat

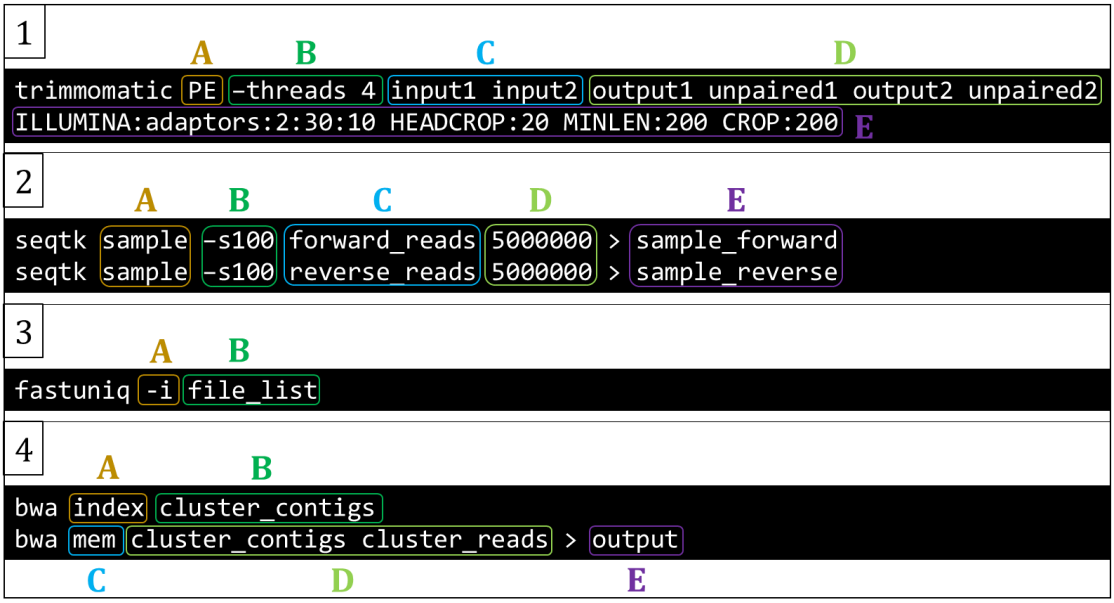
K vykonání nutných úprav sekvenačních dat byl využit software *Trimmomatic* (Bolger et al., 2014). Na čtení byly postupně aplikovány tyto operace – odstranění adaptérových sekvencí (funkce *ILLUMINACLIP*), ořezání prvních 20 nukleotidů (funkce *HEADCROP*), odfiltrování čtení kratších než 200 bp (funkce *MINLEN*) a ořezání čtení na celkovou délku 200 bp (funkce *CROP*). Ořezání začátků a konců čtení je důležité, kvůli horší kvalitě sekvencování na těchto pozicích. Adaptérové sekvence mohou být v různě dlouhých fragmentech začleněny ve výsledných čteních, čímž ovlivňují jejich podobu, a proto je jejich odstranění nezbytné. Konečné ořezání na délku 200 bp je učiněno kvůli sjednocení délky všech čtení. Tato série operací byla aplikována na 2 páry souborů s pair-end sekvenačními daty.

Velikost *FASTQ* souborů byla přibližně 7 GB (soubor s daty B- vzorku) a 23 GB (soubor s daty B+ vzorku). Kvůli jednodušší manipulaci se soubory byl počet čtení v každém z nich zredukován formou pseudonáhodného výběru na 5 milion čtení, přičemž velikost souborů klesla na přibližně 1 GB. K redukci byl použit volně dostupný software *Seqtk*, domovská adresa –https://github.com/lh3/seqtk.

Potenciální duplicitní čtení, která se mohou v souborech nacházet, byla odstraněny pomocí volně dostupného softwaru *FastUniq* (Xu et al., 2012). Šablony použitých příkazů jsou znázorněny na Obr. 7.

3.5 Komparativní analýza na serveru RepeatExplorer

Pro analýzu algoritmem *RepeatExplorer2 clustering* (Novák *et al*., 2010) musí být vstupní soubory upraveny do specifického tvaru. Soubory musí být ve formátu *FASTA*,



Obr. 7 Šablony příkazů využitých při úpravě dat. 1: příkaz pro software *Trimmomatic* A – parametr indikující pair-end čtení na vstupu, B – nastavení výpočetních jader procesoru na 4, C – vstupní pair-end čtení, D – výstupní soubory, output1 a output2 obsahují upravená čtení, unpaired1 a unpaired2 obsahují nespárovaná čtení, E – série úprav, které mají být aplikovány na vstupní soubory. 2: příkaz pro software *seqtk*, musí být aplikován na dva soubory s korespondujícími pair-end čtenmií, A – funkce sample pro náhodné vybrání čtení, B – zvolení tzv. seedu, podle kterého bude náhodný výběr probíhat, u obou souborů musí být totožný, C – vstupní soubory, D – počet čtení, která mají být vybrány, E – výstupní soubory obsahující výběr čtení. 3: příkaz pro software *fastuniq*, A – označení funkce pro odstranění duplicitních čtení, B – soubor obsahující názvy souborů, na které má být algoritmus aplikován. 4: příkazy pro software *BWA*, A – označení funkce pro naindexování kontigů, B – soubor s contigy daného klastru, C – označení funkce pro alignment, D – naindexovaný soubor s kontigy a soubor se čteními tvořícími daný klastr, E – výstupní soubor s výsledky alignmentu.

kde se pod sebou nacházejí sekvence čtení z obou vzorků (forward a reverse z B- vzorku a forward a reverse z B+ vzorku), přičemž se identifikátory čtení ze vzorků musí lišit prefixem (byl zvolen prefix BB pro čtení z B+ vzorku a 00 pro čtení z B- vzorku). Pro potřebné úpravy poskytuje rozhraní Galaxy všechny potřebné nástroje.

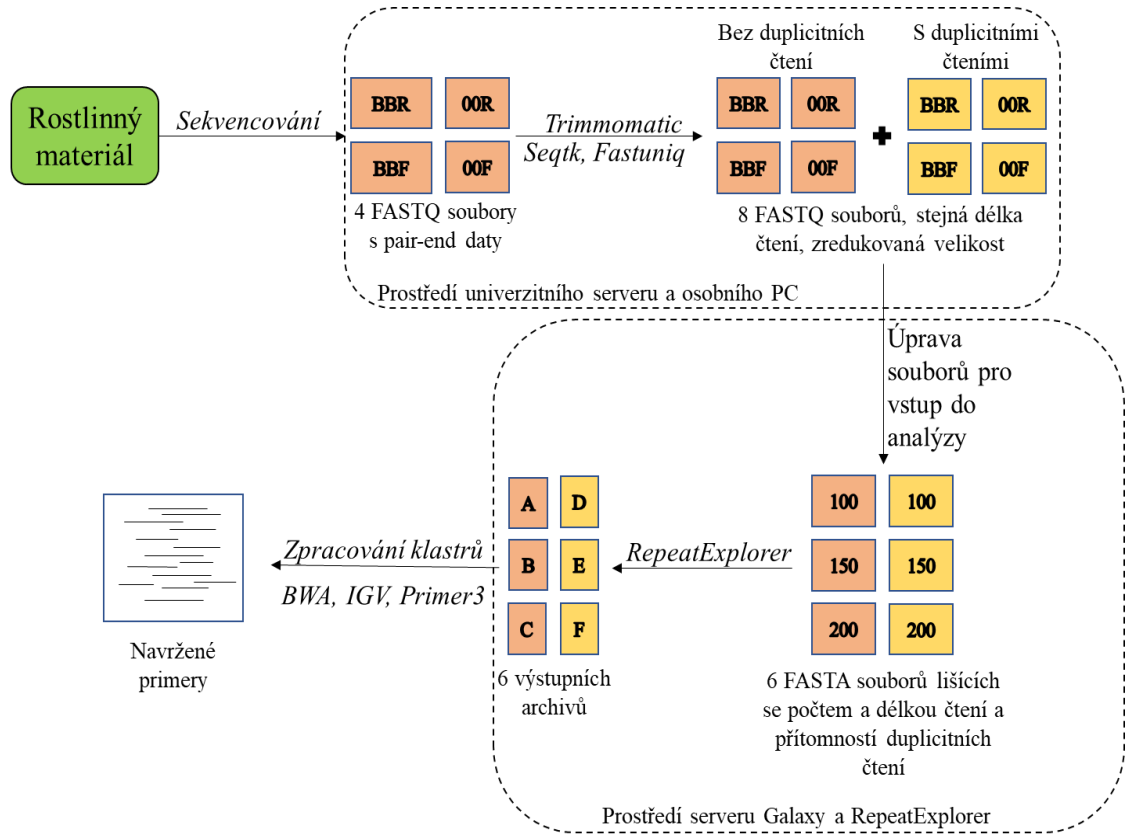
Pro porovnání vlivu délky čtení na výsledek analýzy byly tímto způsobem vytvořeny 3 typy vstupních souborů s rozdílnou délkou a počtem čtení, konkrétně 4 miliony čtení délky 100 bp, 2,8 milionů čtení délky 150 bp a 2 miliony čtení délky 200 bp. Každý soubor obsahoval polovinu pair-end čtení z B+ vzorku a druhou polovinu z B– vzorku. Tyto vstupní soubory byly vytvořeny jak ze souboru s přítomnými duplicitními čteními, tak bez nich, celkově tedy 6 souborů. Při přibližné velikosti genomu *Sorghum purpureosericeum* 2,2 Gb, odpovídají takto zvolené délky a počty čtení přibližnému 8% pokrytí genomu.

Na každý typ vstupních souborů byla aplikována funkce *RepeatExplorer2 clustering* jejíž parametry byly nastaveny tak, aby odpovídaly podobě vstupních souborů (pair-end soubory, dvoumístné označení čtení apod.), databází proteinových domén byla zvolena *REXdb Viridiplantae v.3.0*, přičemž pro vyhledání v těchto doménách byl zvolen algoritmus *blastx* s třípísmennou délkou slova a velikostní limit pro anotaci klastrů byl nastaven na 0,01 %

3.6 Charakteristika repetitivních sekvencí a navržení primerů

K charakterizaci repetitivních sekvencí ve vzorku a k identifikaci B-specifických klastrů (alespoň z 80 % tvořeny B+ čteními), byl využit software napsaný v programovacím jazyce Python 3.7 (podrobně popsaný v kapitole Výsledky) aplikovaný na výstupní soubory komparativní analýzy. Tímto způsobem byly také extrahovány B+ čtení, která B-specifické klastry tvoří a také sekvence contigů (konsenzuální sekvence vzniklé překryvem čtení při alignmentu), které byly algoritmem *RepeatExplorer2 clustering* sestaveny. Mezi contigy a B+ čteními byl vytvořen alignment (porovnání a seřazení čtení tak, aby jejich sekvence odpovídaly contigům) pomocí softwaru *BWA* verze 0.7.17 (Li a Durbin, 2017, šablona na Obr. 1). Poté byly na základě vizualizace výsledků alignmentu softwarem *IGV* (Robinson *et al*., 2011) vybrány úseky contigů s největším pozičním pokrytím. Z těchto sekvencí byly softwarem *Primer3* (Koressaar a Remm, 2007) navrženy primery.

Celkový průběh analýzy a využitý software vedoucí k navržení primerů je znázorněn na Obr. 8.



Obr. 8 Schéma znázorňující jednotlivé kroky zpracování genomických dat až po navržení primerů. Zkratky BBR, BBF, 00R a 00F reprezentují jednotlivé páry (reverse – R, forward – F) pair-end čtení ze vzorku s B chromozomy (BB) a čtení ze vzorku bez B chromozomů (00). Hodnoty 100, 150 a 200 označují délky čtení.

4 VÝSLEDKY A DISKUZE

4.1 Výsledek algoritmu RepeatExplorer2 clustering

Implementace algoritmu *RepeatExplorer2 clustering* je založena na sjednocení algoritmů *RepeatExplorer2* a *TAREAN*, díky tomu je výsledkem aplikace algoritmu klastrová analýza charakterizující jak repetitivní sekvence ve vstupním souboru spolu se zastoupením jednotlivých typů čtení, tak určení sekvencí potencionálních tandemových satelitů. Výsledky jsou uloženy ve dvou typech souborů – komprimovaném *ZIP* souboru a *HTML* souboru.

HTML soubor umožňuje grafické prohlížení výsledků ve webovém prostředí. V rozhraní lze skrze hypertextové odkazy procházet rozdílná výsledná data algoritmu. Hlavní stránka obsahuje základní kvantitativní údaje týkající se průběhu analýzy (počet vstupních čtení, počet analyzovaných čtení apod), vizualizace struktury klastrů a odkazy na konkrtétní výsledky – *Tandem repeat analysis* (výsledky algoritmu *TAREAN*), *Cluster annotation* (anotace klastrů vzniklých klastrovou analýzou algoritmem *RepeatExplorer2*), *Supercluster annotation* (anotace superklastrů – seskupení klastrů majících společné čtení) a *Repeat annotation summary* (shrnutí anotace repetitivních úseků). Veškeré informace nacházející se v těchto položkách jsou pouze vizualizací, nebo shrnutím datových souborů obsažených v *ZIP* archivu, proto je jejich popis zahrnut až v charakteristice těchto datových souborů. Archiv rovněž obsahuje *HTML* soubory, které odkazují na jednotlivé části hlavního *HTML* souboru.

Komprimovaný *ZIP* soubor je strukturován do velkého množství adresářů a souborů, proto budu podrobněji charakterizovat pouze soubory, jež byly dále analyzovány v rámci zpracování výlsedků, nebo odkazují na vizualizace výsledků. V hlavním adresáři to jsou tyto soubory:

1. *COMPARATIVE_ANALYSIS_COUNTS.csv* – soubor s údaji o počtech B- a B+ čtení v jednotlivých klastrech, spolu se zařazením klastrů do příslušných superklastrů. Jsou zde zahrnuty všechny vytvořené klastry, nejen anotované. Jednotlivé hodnoty jsou odděleny tabulátory.
2. *CLUSTER_TABLE.csv* – soubor s podrobnější charakterizací anotovaných klastrů (počet čtení v klastru, anotace proteinových domén, TAREAN anotace) a kvantitativních údajích proběhlé analýzy (počet klastrů, singletů, superklastrů,

čtení v klastrech a analyzovaných čteních). Jednotlivé hodnoty jsou odděleny tabulátory.

3. *TAREAN_consensus_rank_x.fasta – FASTA* soubory s identifikovanými tandemovými repetitivními sekvencemi. Na základě čísla *x* obsahují – satelitní sekvence identifikované s vysokou spolehlivostí (x = 1), satelitní sekvence identifikované s nízkou spolehlivostí (x = 2), domnělé LTR elementy (x = 3) a rDNA (x = 4).
4. *index.html –HTML* totožný s výstupním *HTML* souborem.
5. *summarized_annotation.html* – odkaz na shrnutí výsledků anotace repetic, obsahuje počty čtení, klastrů, superklastrů a poměrné zastoupení jednotlivých typů repetitivních elementů ve vzorku.
6. *supercluster_report.html* – odkaz na shrnutí výsledků anotací k jednotlivým superklastrům, obsahuje poměrné zastoupení a počet čtení vztahujících se k jednotlivým typům repetitivních elementů ve vzorku.
7. *tarean_report.html* – obsahuje sekvence identifikovaných tandemových repetic, totožné informace co se nacházejí v souborech *TAREAN_consensus_rank_x.fasta.*
8. *documentation.html* – soubor s přehledem a popisem veškerých veličin sloužících k popisu výsledků analýzy.

V adresářích obsahujících údaje o jednotlivých anotovaných klastrech (*/seqclust/clustering/clusters/dir_CLxxxx/*, kde za x je dosazen identifikátor klastru spolu s předcházejícím počtem 0, tak by byla sekvence čísel čtyřmístná) se jedná o tyto soubory:

1. *protein_database_annotation.csv* – soubor se čteními, která byla identifikována jako součást určité proteinové domény. Obsahuje identifikátor čtení a anotaci daného elementu spolu s dalšími veličinami. Jednotlivé hodnoty jsou odděleny tabulátory.
2. *reads.fasta – FASTA* soubor se sekvencemi čtení, která tvoří daný klastr.
3. *contigs.fasta – FASTA* soubor se sekvencimi poskládaných contigů daného klastru.

4.1.1 Systém anotací čtení a klastrů

Při využití databáze *REXdb Viridiplantae v. 3.0* k identifikaci proteinových domén transpozibilních elementů dochází k porovnání každého analyzovaného čtení se sekvencemi v této databázi a v případě dostatečné podobnosti, je čtení anotováno v tomto tímto způsobem: */cesta/ve/stromu/klasifikace:rodina-doména*. Rodina v tomto případě označuje rodinu repetitivních elementů, doména je ve zkratkovitém tvaru, např. *INT* pro integrasu. Anotací proteinové domény je např. *Class_I/LTR/Ty1_copia/SIRE:Ty1-INT*.

V případě klastrů existují dva typy anotací – automatické anotace repetitivních elementů a *TAREAN* anotace. Automatická anotace klastru může být určena několika způsoby. Jedním z nich je vybrání repetitivního elementu, který převažuje u anotací čtení tvořících superklastr, jehož je daný klastr součástí (anotace je poté ve tvaru *All/repeat/mobile_element/klasifikace*). Dále může být u klastru odhalena podobnost s rDNA repeticí (anotace ve tvaru *All/repeat/rDNA/typ_rDNA*), nebo klastr přímo reprezentuje tandemovou repetici (anotace ve tvaru *All/repeat/satellite*). Pokud nelze na základě žádného z předchozích postupů anotaci vytvořit, je klastr anotován jako All. Anotace také může být v neúplném tvaru, např. *ALL/repeat/mobile_element*, v takovém případě nedošlo k přesnější anotaci.

TAREAN anotace musí být v jednom z těchto tvarů – *Putative satellites (high confidence)* pro satelitní sekvence identifikované s vysokou pravděpodobností správně, *Putative satellites (low confidence)* pro satelitní sekvence identifikované s nízkou pravděpodobností správně, *Putative LTR elements* pro identifikované *LTR* elementy, *rDNA* pro identifikované repetice typické pro *rDNA* a *Other* pro ostatní klastry, u kterých nebyly odhaleny žádné tandemově repetitivní úseky.

4.2 Software k vizualizaci a zpracování výsledků komparativní analýzy

K automatizovanému zpracování výsledků komparativní analýzy dvou vzorků pomocí algoritmu *RepeatExplorer2 clustering* a vizualizaci výsledků klastrové analýzy byl implementován software v programovacím jazyce Python 3.7 (Obr. 1). Implementace je postavena na základních prvcích jazyka a několika instalovatelných knihoven – *pandas* s nástroji pro práci a manipulaci se soubory ve formátu *CSV*, *tkinter* nástroji pro tvorbu grafického uživatelského rozhraní, *matplotlib* s nástroji pro tvorbu vizualizací, *numpy*

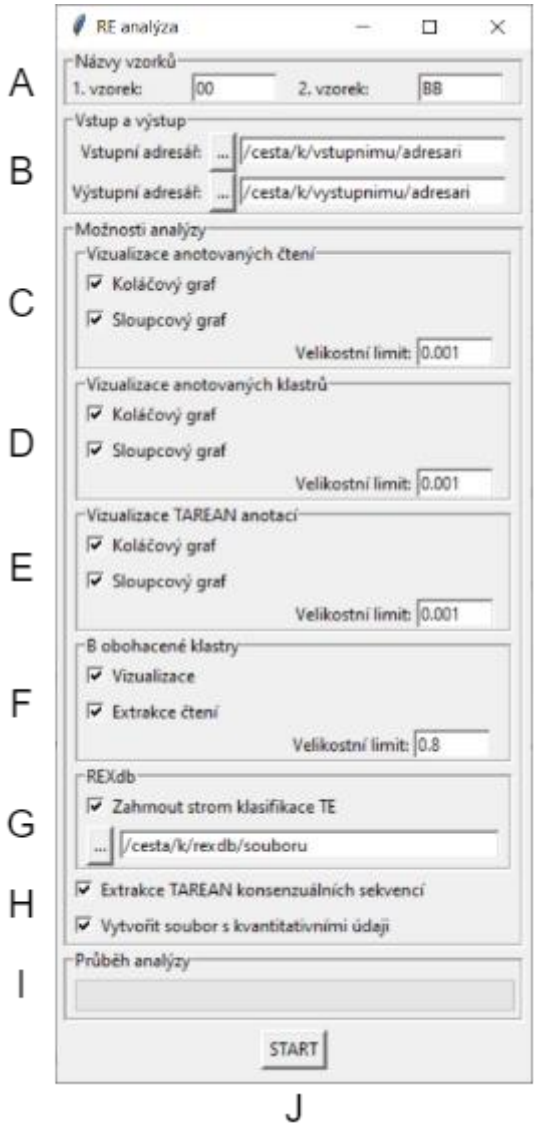
s nástroji pro práci se sofistikovanějšími číselnými datovými typy, *anytree* s nástroji pro stromové vizualizace a *shutil* s nástroji pro manipulaci se soubory.

Jediným vstupním souborem softwaru je archiv vytvořený algoritmem *RepeatExplorer2 clustering* v extrahované podobě. Jako databáze proteinových domén, dle které jsou klastry a čtení anotovány, musí být zvolena *REXdb Viridiplantae v. 3.0*.

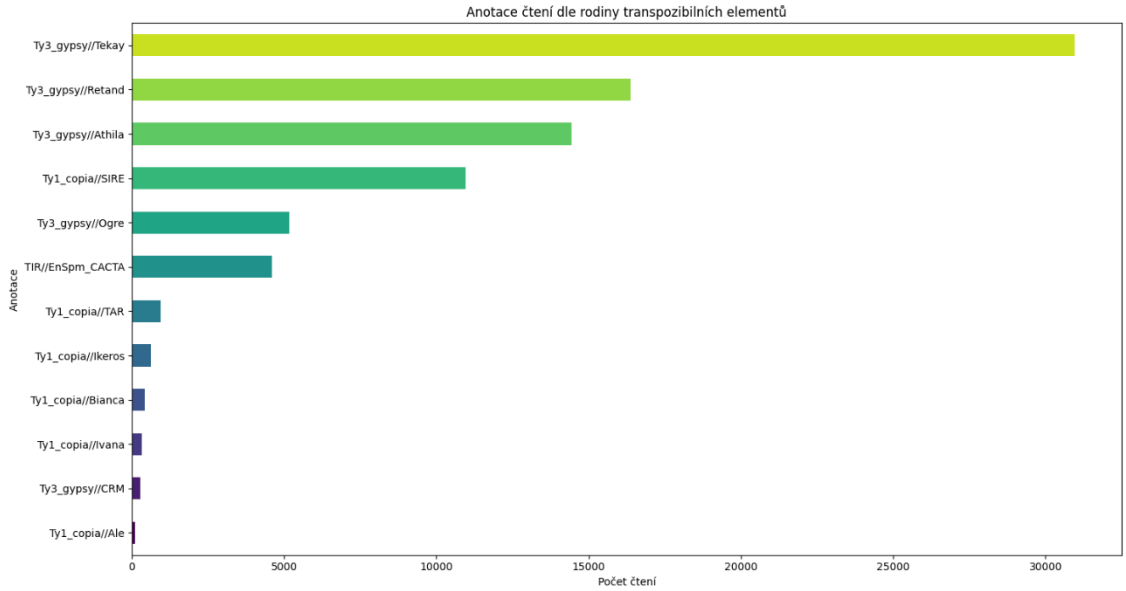
Software je po instalaci spustitelný pomocí souboru *main.exe*.

4.2.1 Funkce softwaru

V následujícím seznamu jsou popsány veškeré funkce vytvořeného softwaru. K jejich správnému fungovají je nutné správně navolit cestu k extrahovanému archivu (výstup



Obr. 9 Rozhraní vyvinutého softwaru ke zpracování výsledků algoritmu *RepeatExplorer2 clustering.*

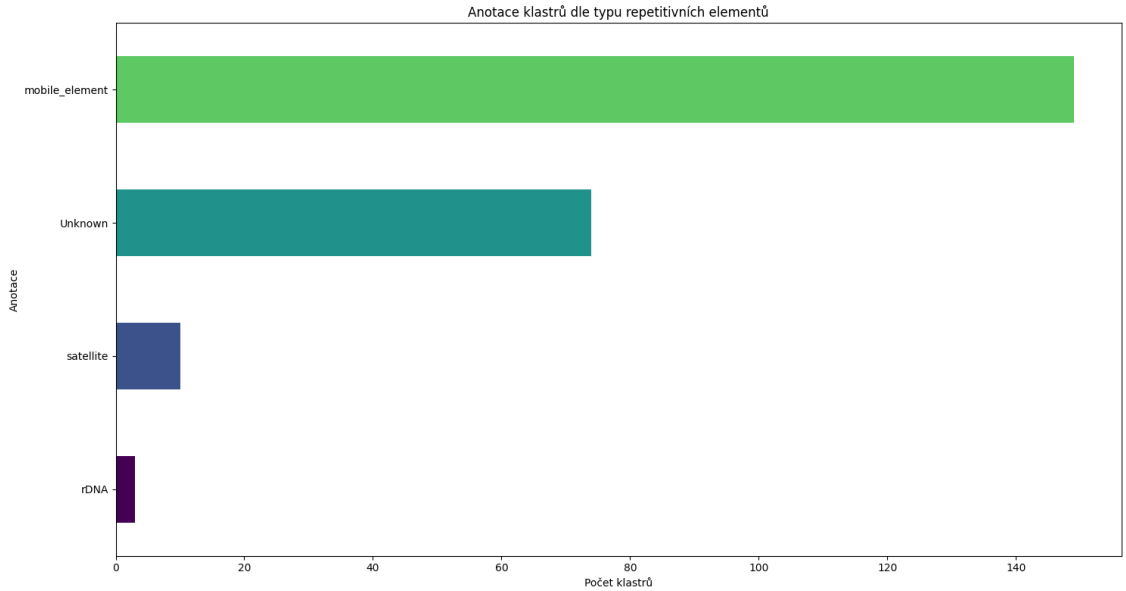


Obr 10. Výstupní soubor funkce *Vizualizace anotovaných čtení* na data vzniklá aplikací algoritmu *RepeatExplorer2 clustering* na vstupní soubory bez duplicitních čtení a délkou čtení 100 bp. Data jsou rozdělena dle typu rodiny transposibilních elementů, u názvů anotací se před názvem rodiny elementů nachází jejich zařazení do řádu.

algoritmu *RepeatExplorer2 clustering*) a k adresáři, ve kterém se budou nacházet výsledky (sekce Vstup a výstup, obr. 9B). Rovněž je nutné nadefinovat prefixy použité pro označení vzorku při komparativní analýze, prefixy musí odpovídat názvům sloupců v souboru *COMPARATIVE_ANALYSIS_COUNTS.csv* (sekce Názvy vzorků, obr. 9A). V popisu funkcí jsou využity tyto symboly – L pro velikostní limit, C pro celkový počet čtení, která tvoří klastry, c pro počet čtení v konkrétním klastru, K pro celkový počet anotovaných klastrů a dvojice znaků ⌈ ⌉ pro zaokrouhlení nahoru.

1. *Vizualizace anotovaných čtení* – funkce (Obr. 9C), která na základě analýzy souborů *protein_database_annotation.csv* příslušejícím jednotlivým anotovaným klastrům vytvoří dle volby uživatele koláčové (výstupní soubory v adresáři /pie_plots/reads/), nebo sloupcové grafy (výstupní soubory v adresáři /bar_plots/reads/) znázorňující zastoupení jednotlivých druhů proteinových anotací mezi čteními, jejichž počet je alespoň $\lceil L \times C \rceil$. Názvy výstupních souborů mají prefix *reads_*. Příklad vizualizace na Obr. 10.

2. *Vizualizace anotovaných klastrů* – funkce (Obr. 9D), která na základě analýzy souboru *CLUSTER_TABLE.csv* vytvoří dle volby uživatele koláčové (výstupní soubory v adresáři /pie_plots/clusters/), nebo sloupcové grafy (výstupní soubory v aresáři /bar_plots/clusters/) znázorňující zastoupení jednotlivých druhů

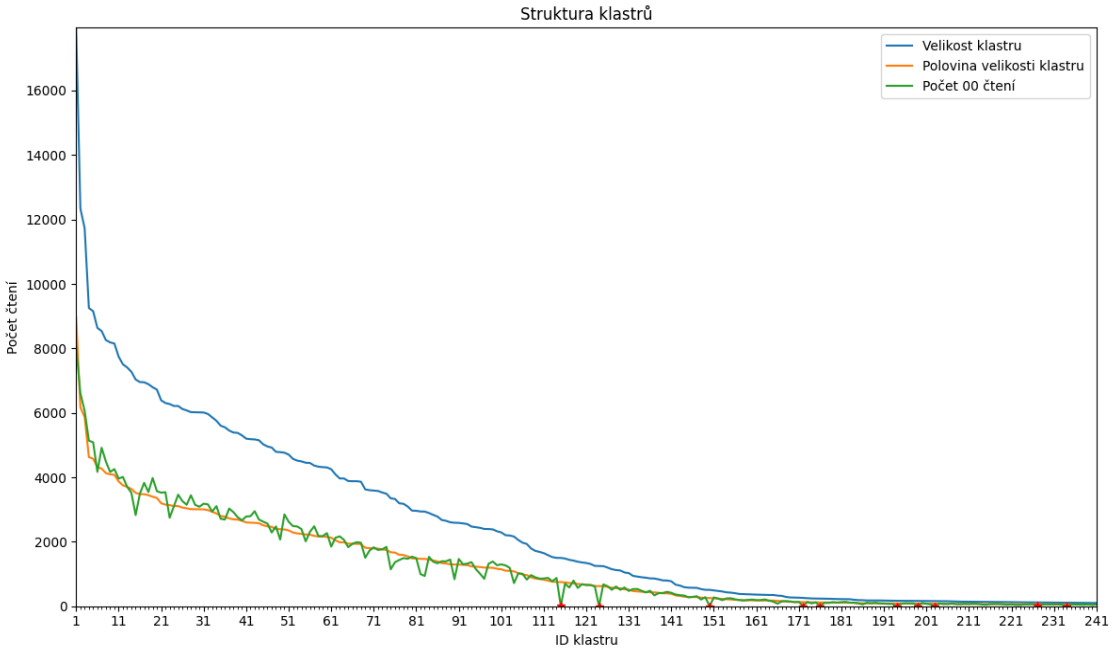


Obr. 11 Výstupní soubor funkce *Vizualizace anotovaných klastrů* na data vzniklá aplikací algoritmu *RepeatExplorer2 clustering* na vstupní soubory bez duplicitních čtení a délkou čtení 100 bp. Data jsou rozdělena dle typu repetice.

proteinových anotací mezi klastry, jejichž počet je alespoň $\lceil L \times \mathrm{K} \rceil$. Názvy výstupních souborů mají prefix *clusters_*. Příklad vizualizace na Obr. 11.

3. *Vizualizace TAREAN anotací* – funkce (Obr. 9E), která na základě analýzy souboru *CLUSTER_TABLE.csv* vytvoří dle volby uživatele koláčové (výstupní soubory v adresáři /pie_plots/tarean/), nebo sloupcové grafy (výstupní soubory v adresáři /bar_plots/tarean/) znázorňující zastoupení *TAREAN* anotací mezi anotovanými klastry, jejichž počet je alespoň $\lceil L \times K \rceil$. Výstupní soubory s grafy mají prefix *tarean_*.

4. *B obohacené klastry* – funkce (Obr. 9F), která při označení volby *Vizualizace* vytvoří graf znázorňující poměr zastoupení B+ a B- čtení v jednotlivých klastrech spolu s označením klastrů, u nichž zastoupení B+ čtení přesahuje $\lceil L \times \mathrm{c} \rceil$. Graf je uložen v souboru *cluster_overview.png*. Při označení volby *Extrakce čtení* dojde k vytvoření adresáře */bb_enriched/* s podadresáři korespondující s názvy klastrů, u nichž je počet B+ čtení alespoň $\lceil L \times \mathrm{c} \rceil$. Tyto podadresáře obsahují soubor *reads.fasta* se sekvencemi B+ čtení, která tvoří daný klastr a soubor *contigs.fasta* se sekvencemi sestavených kontigů daného klastru. Příklad vizualizace na Obr. 12.



Obr. 12. Vizualizace struktury anotovaných klastrů vzniklých aplikací algoritmu *RepeatExplorer2 clustering* na vstupní soubor se čteními délky 100 bp a přítomností duplikacitních čtení. Červené hvězdy označují klastry, jež jsou alespoň z 80 % tvořeny B čteními.

5. *Extrakce TAREAN konsenzuálních sekvencí* – funkce (Obr. 9F) pro vytvoření *FASTA* souborů se satelitními sekvencemi identifikovanými algoritmem *TAREAN*. Soubory se nacházejí v asresáři */tarean_sequences/* a jejich jména odpovídají anotací sekvencí algoritmem *TAREAN*.

6. *REXdb* – funkce (Obr. 9G) pro vygenerování textové podoby klasifikačního stromu transpozibilních elementů *classification.txt*, na základě zpracování *FASTA* souboru s klasifikacemi použité REXdb. Klasifikace koresponduje s anotacemi použitými v grafech.

7. *Vytvořit soubor s kvantitativními údaji* – funkce (Obr. 9H) pro vygenerování souboru *quantitative_data.csv* se základními kvantitativními údaji z analýzy, mezi které patří počty čtení v klastrech, klastrů, superklastrů, singletů, analyzovaných čtení a anotovaných klastrů aj. Velikostní limit lze ke každé funkci, kde je vyžadován libovolně nastavit. Veškeré výstupní soubory jsou uloženy ve dvou typech souborů – *PNG* soubor s vizualizací daných vlastností a *CSV* soubor s kvantitativními údaji, na základě kterých byla vizualizace zkonstruována, názvy těchto souborů se liší pouze příponou charakterizující typ souboru. Jediným výstupním souborem, k jehož vytvoření dojde při jakékoli konfiguraci softwaru je soubor *log.txt* obsahující popis daného běhu softwaru.

Průběh analýzy je reprezentován pomocí stavového řádku ve spodní části rozhraní (obr. 9I). Práce softwaru je spuštěna kliknutím na tlačítko START (Obr. 9J)

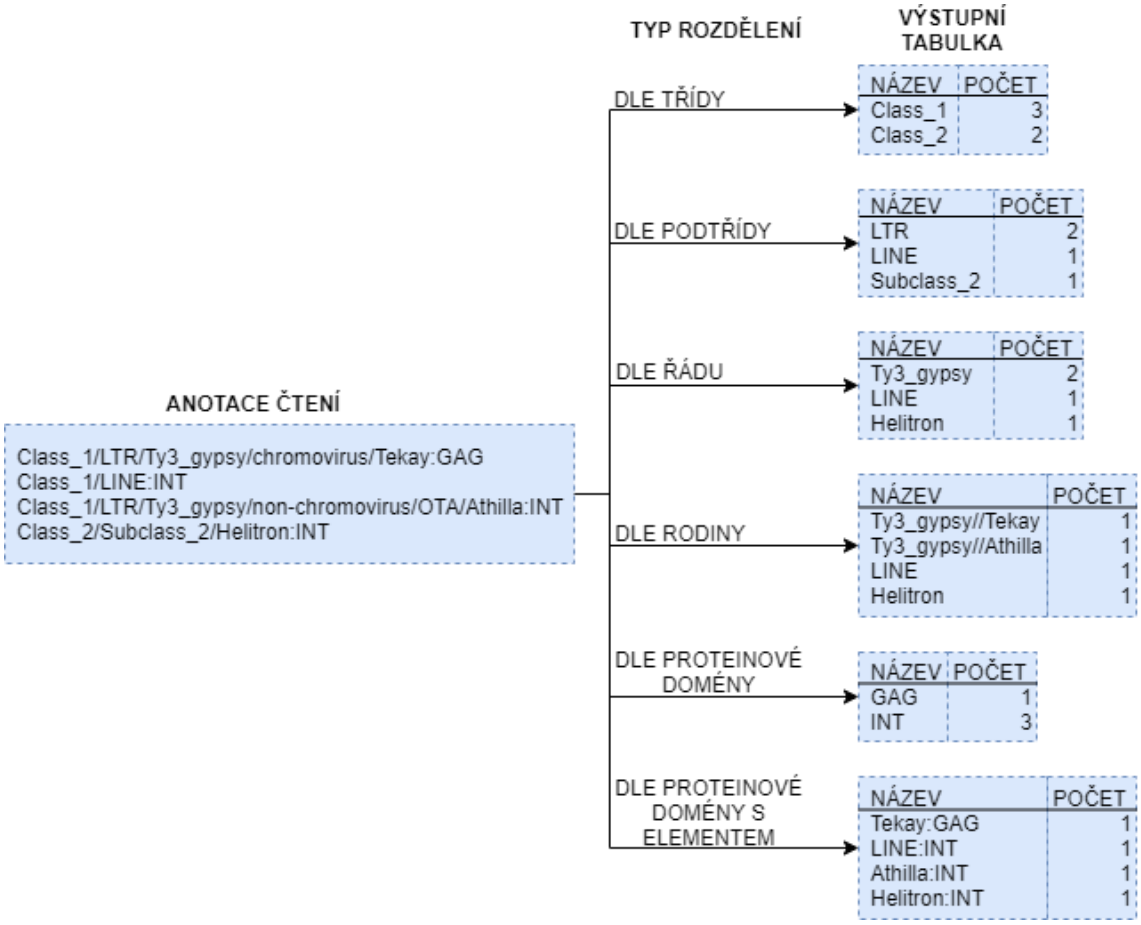
4.2.2 Způsoby vizualizací anotací

Výstupní soubory s grafovými vizualizacemi se odlišují podle části anotace, dle které jsou data ve vizualizacích rozdělena. Názvy výstupních souborů se dle typu odlišují příponami.

V případě anotace čtení se jedná o tyto přípony:

- *_class* – rozdělení dle třídy transposibilních elementů, viz obr.
- *_subclass* – rozdělení dle podtřídy transposibilních elementů
- *_family* – rozdělení dle rodiny transposibilních elementů
- *_order* – rozdělení dle řádu transposibilních elementů
- *_protein_domain* – rozdělení dle proteinové domény transposibilních elementů
- *_protein_domain_with_family* – rozdělení dle proteinové domény spolu s rodinou transposibilních elementů

Rozdělení anotací na třídu, podtřídu, řád a rodinu je určeno kvůli přehlednosti, jednotlivé zařazení ale nemusí reflektovat skutečnou klasifikaci daného elementu. Jména jednotlivých částí anotací byla odvozena od jejich podrobnosti, přičemž klasifikační strom (Příloha 8.4) odpovídá anotací transpozibilních elementů dle databáze *REXdb version 3*. V případě, že anotace čtení nedosahuje úrovně, dle které má být anotace rozdělena, je vždy využita nejpodrobnější část anotace. Při rozdělení dle rodiny elementu je u podrobných anotací přiřazen také řád, do kterého daná rodina patří. Mechanismus rozdělení a vizualizace anotací je naznačen v Obr. 13.



Obr 13 Ilustrace tabulek, ze kterých vycházejí vizualizace anotovaných čtení. Na základě takových tabulek je poté vytvořen graf (koláčový, sloupcový).

Vizualizace anotovaných klastrů obsahují rovněž rozdělení dle klasifikace mobilních elementů, ale navíc je přidáno rozdělení dle typu repetitivního elementu (přípona souboru *_repeat_type*). U *TAREAN* anotací nedochází k žádnému rozdělování anotací.

Informace vizualizovaná jak v podobě koláčových grafů, tak v podobě sloupcových grafů je totožná, odlišuje se pouze v podobě vizualizace. Barvy použité pro výřezy koláčového grafu a sloupce sloupcového grafu označující stejný druh anotace mají rovněž stejné barevné označení.

4.2.3 Výsledky aplikace algoritmu *RepeatExplorer2 clustering*

Aplikací algoritmu *RepeatExplorer2 clustering* na 6 různých vstupních souborů specifikovaných v kapitole 3.4, vzniklo 6 výstupních archivů. Na extrahovaný obsah každého z těchto archivů byly poté aplikovány všechny funkce implementovaného softwaru, velikostní limity pro vizualizace anotovaných prvků byly nastaveny na 0,001, velikostní limit pro B obohacené klastry byl nastaven na 0,8. V následujících

Tab. 1 Kvantitativní údaje výsledků klastrové analýzy algoritmem *RepeatExplorer2 clustering* na 3 vstupní soubory s přítomností duplicitních čtení.

Délka čtení	100	150	200
Celkový počet čtení	4000000	2800000	2000000
Počet analyzovaných čtení	959240	910055	894787
Počet identifikovaných klastrů	28230	28706	29463
Počet superklastrů	28082	28601	29395
Počet singletů	273717	256566	250815
Počet čtení v klastrech	685523	653489	643972
Počet anotovaných klastrů	241	239	224
Minimální počet čtení tvořících klastr	101	94	90
Maximální počet čtení tvořících klastr	17950	15840	14662
Průměrný počet čtení tvořících klastry	2451	2343	2463

podkapitolách jsou získané informace popsány a v rámci různých vstupních souborů porovnány.

4.2.4 Číselná charakteristika klastrové analýzy

Počet analyzovaných čtení (čtení, jež byla zahrnuta v klastrové analýze) se lišil v souvislosti s velikostí vstupního souboru, nicméně tento rozdíl neodpovídal rozdílu ve velikostech vstupních souborů (přesná data v Tab. 1 a Tab. 2). V případě vstupních dat s duplicitními čteními se poměr čtení využitých v analýze pohyboval mezi 44 % (u čtení délky 200 bp,) a 24 % (u čtení délky 100 bp). U vstupních souborů s absencí duplicitních čtení byl poměr analyzovaných čtení přibližně o 1 % vyšší, tato skutečnost by mohla souviset s duplicitními čteními v klastru, díky kterým dojde k rychlejší konstrukci klastru, bez nutnosti většího počtu čtení v klastru.

Tab. 2 Kvantitativní údaje výsledků klastrové analýzy algoritmem *RepeatExplorer2 clustering* na 3 vstupní soubory s absencí duplicitních čtení.

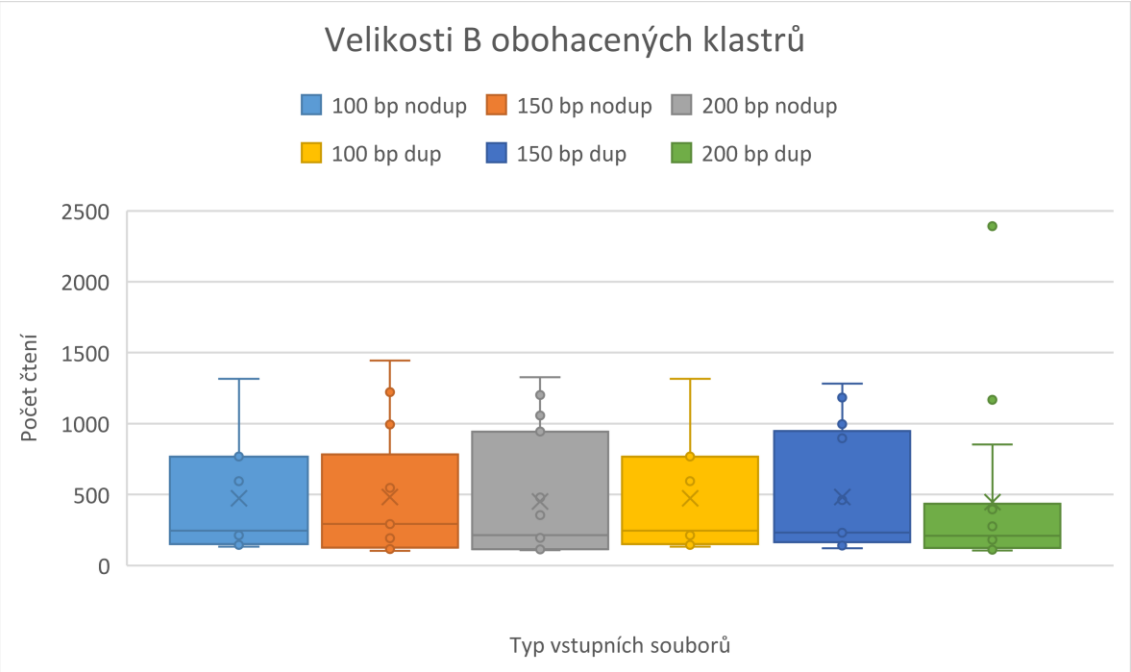
Délka čtení	100	150	200
Celkový počet čtení	4000000	2800000	2000000
Počet analyzovaných čtení	989305	933378	913221
Počet klastrů	29278	29499	29833
Počet superklastrů	29122	29392	29759
Počet singletů	281193	261594	255406
Počet čtení v klastrech	708112	671784	657815
Počet anotovaných klastrů	236	241	227
Minimální velikost anotovaného klastru	98	93	91
Maximální velikost anotovaného klastru	18449	16570	15146
Průměrná velikost anotovaného klastru	2581	2394	2489

Nejvíce klastrů bylo identifikováno při analýze souboru s nejdelšími čteními (200 bp), nicméně v tomto případě také dochází k nejmenšímu počtu anotovaných klastrů (tzn. klastrů obsahujících alespoň 0.01 % z celkového počtu analyzovaných čtení). Avšak obecná korelace mezi délkou čtení a počtem identifikovaných klastrů nelze z výsledku určit, protože v případě dat s duplikacemi bylo nejvíce klastrů anotováno u analyzovaného vstupního souboru se 100bp čteními, ale u dat bez duplikací se jednalo o soubor se 150bp čteními.

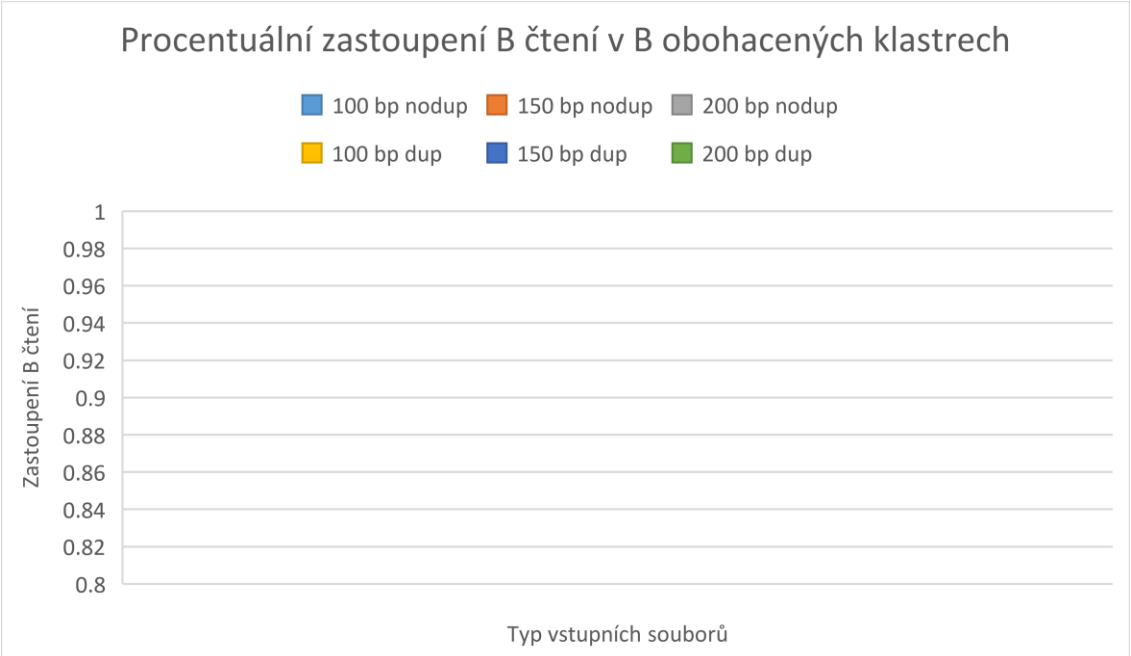
Navzdory odlišnému počtu analyzovaných čtení je poměr čtení v klastrech u všech typů vstupních souborů téměř totožný, přibližně 72 %. Pro klastry identifikované v souboru dat s absencí duplicitních čtení, je typický větší počet čtení v největším klastru, nicméně průměrná velikost anotovaných klastrů je velmi podobná ve všech případech.

4.2.5 Charakteristika B-specifických klastrů

Přítomnost duplicitních čtení neměla zásadní vliv na počet B-specifických klastrů mezi anotovanými klastry, na rozdíl od délky čtení, která s počtem B-specifických klastrů přímo korespondovala (s délkou čtení rostl i počet B-specifických klastrů). I když analýzou souboru se čteními délky 200 bp došlo k vytvoření největšího množství B-specifických klastrů, dosahovaly nejmenší průměrné velikosti (viz Obr. 14).



Obr. 14 Porovnání velikostí B obohacených klastrů mezi jednotlivými vstupními soubory. Zkratka dup označuje soubor s duplikacemi, nodup naopak absenci duplikací.



Obr. 15 Porovnání zastoupení B čtení v B obohacených klastrech mezi jednotlivými vstupními soubory. Zkratka dup označuje soubor s duplikacemi, nodup naopak absenci duplikací.

Zastoupení B čtení mezi B-specifickými klastry se u všech typů souborů průměrně pohybovalo v rozmezí 96 až 98 % a ve všech případech souborů se nacházely klastry tvořené pouze B+ čteními (viz Obr. 15).

4.2.6 Charakteristika anotovaných čtení

Při analýze vstupních dat s přítomností duplicitních čtení bylo anotováno průměrně 91606 čtení, oproti tomu v případě vstupních dat bez duplicitních čtení se jednalo o 94211 čtení. Rovněž platí, že se vzrůstající délkou čtení roste i počet anotovaných čtení, navzdory tomu, že počet analyzovaných čtení roste s jejich klesající délkou. Poměr zastoupení DNA transpozónů vůči retrotranspozónům je u všech souborů téměř totožný, 95 % ve prospěch retrotranspozónů.

Tab. 3 Počty anotovaných transpozibilních elementů u čtení. Anotace vznikly aplikací algoritmu *RepeatExplorer2* na jednotlivé typy vstupních souborů. Jedná se o výběr několika proteinových domén, tak aby byla naznačena povaha počtů, tabulka v plném rozsahu je v Příloze 2.

Typ vstupu	S duplikacemi			Bez duplikací		
Délka čtení	100	150	200	100	150	200
Tekay	29931	31913	34166	30951	33102	34890
Retand	15618	16867	17497	16365	17165	18109
Athila	13935	16118	17994	14438	16706	18093
SIRE	10733	11907	13181	10952	12332	13433
Ogre	5228	5497	5707	5160	5616	5922
Ale	1	175	360	96	147	415
Angela	0	4	141	0	13	140
Helitron	0	1	127	0	2	100

Ve většině případů se počet anotovaných elementů u jednotlivých vstupních souborů lišil úměrně s počtem anotovaných čtení, nicméně u několika elementů došlo k velkému počtu anotaci (elementy z řádu *Helitron* a *Angela* z řádu *Ty1/copia*), nebo k více jak dvojnásobnému počtu anotací (element *Ale* z řádu *Ty1/copia*) pouze u vstupních souborů s délkou čtení 200 bp, viz Tab. 3. Podobný trend lze vidět u anotovaných proteinových domén, kde jsou odlišné počty domén typické pro specificky zastoupené elementy (proteinová doména *Helitron-HEL1*, *Helitron-HEL2* je výrazně více zastoupena ve vstupním souboru s délkou čtení 200 bp s přítomností duplicit, v případě absence duplicit je anotována hlavně doména *Helitron-HEL2*), což odkazuje na správnost získaných dat (zastoupení elementů koresponduje se zastoupením proteinových doméne těchto elementů). Mezi nejčastěji anotované rodiny elementů patří *Tekay, Retand, Athila* z řádu *Ty3-gypsy* a *SIRE* z řádu *Ty1/copia* (dohromady tvoří více jak 50 % anotací). Nejčastěji anotovanými proteinovými doménami jsou domény *INT, RT, GAG* a *RH* patřící k řádu

Tab. 4 Počty anotovaných proteinových domén transpozibilních elementů u čtení. Anotace vznikly aplikací algoritmu *RepeatExplorer2* na jednotlivé typy vstupních souborů. Jedná se o výběr několika proteinových domén, tak aby byla naznačena povaha počtů, tabulka v plném rozsahu je v Příloze 1.

Typ vstupu	S duplikacemi			Bez duplikací		
Délka čtení	100	150	200	100	150	200
Ty3-INT	20851	21588	21528	21911	21977	22180
Ty3-RT	12581	12922	13953	12984	13659	14130
Ty3-GAG	10910	12438	13792	11073	12964	14114
Ty3-RH	9973	10669	11338	10311	10971	11795
Ty3-PROT	5592	7513	9197	5669	7564	9175
Helitron-HEL1	0	1	52	0	1	7
Helitron-HEL2	0	0	75	0	1	93

Tab. 5 Počty anotovaných klastrů dle druhu repetitivních sekvencí. Anotace vznikly aplikací algoritmu *RepeatExplorer2 clustering* na jednotlivé typy vstupních souborů.

Typ vstupu	S duplicitami			Bez duplicit		
Délka čtení	100	150	200	100	150	200
mobile_element	151	152	142	149	157	149
Unknown	77	75	70	74	71	66
satellite	10	8	8	10	8	7
rDNA	3	3	3	3	3	3
plastid	0	1	1	0	2	2

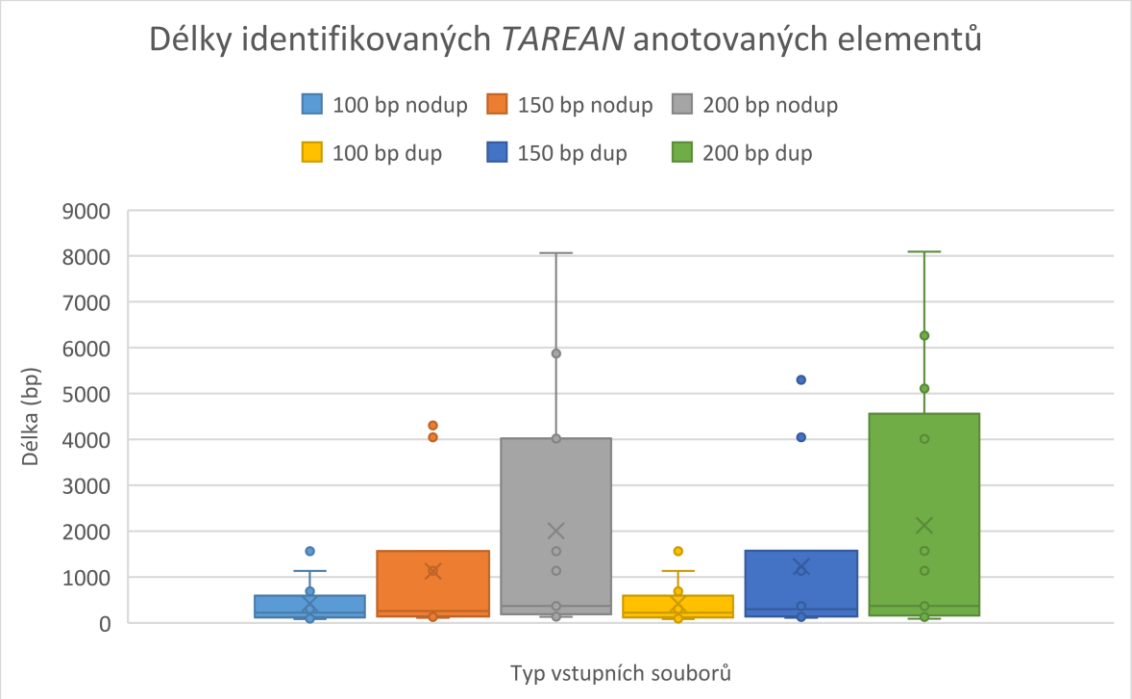
Ty3/gypsy Počty některých anotovaných proteinových domén transpozibilních elementů jsou zobrazeny v Tab. 4.

4.2.7 Charakteristika anotovaných klastrů

Nejčastěji byly klastry anotovány jako transpozibilní elementy (v rozmezí 62 až 65 % z celkového počtu anotovaných klastrů). Nicméně u velkého množství klastrů (v rozmezí 28 až 32 % z celkového počtu anotovaných klastrů) nebyl žádný repetitivní element v klastru identifikován, a proto jejich anotace nebyla určena. Mezi nejčastěji anotované konkrétní repetitivní elementy patřily *Athila, Tekay, Retand* a *SIRE*, což koresponduje s nejčastějšími anotacemi čtení. Rovněž byly některé klastry anotovány jako satelitní sekvence, rDNA aj. Přehled počtů anotovaných klastrů dle typu repetitivní sekvence je zobrazen v Tab. 5. Počty nenaznačují žádný výraznější vliv typu a délky čtení na druhy anotací. Počty anotovaných konkrétních typů repetitivních sekvencí jsou zobrazeny částečně v Tab. 6.

Tab. 6 Počty anotovaných klastrů dle konkrétního typu repetitivních sekvencí. Anotace vznikly aplikací algoritmu *RepeatExplorer2 clustering* na jednotlivé typy vstupních souborů. Jedná se o výběr několika anotací, tak aby byla naznačena povaha počtů, tabulka v plném rozsahu je v Příloze 3.

Typ vstupu	S duplicitami			Bez duplicit		
Délka čtení	100	150	200	100	150	200
Unknown	77	74	69	74	70	65
Athila	35	34	32	36	35	30
Tekay	35	37	36	36	38	35
Retand	23	20	17	19	20	15
SIRE	15	15	12	16	17	13
EnSpm_CACTA	14	11	10	13	14	11
satellite	10	8	8	10	8	7
LTR	7	6	5	8	9	8



Obr. 16 Porovnání délek identifikovaných *TAREAN* element. Zkratka dup označuje soubor s duplikacemi, nodup naopak absenci duplikací.

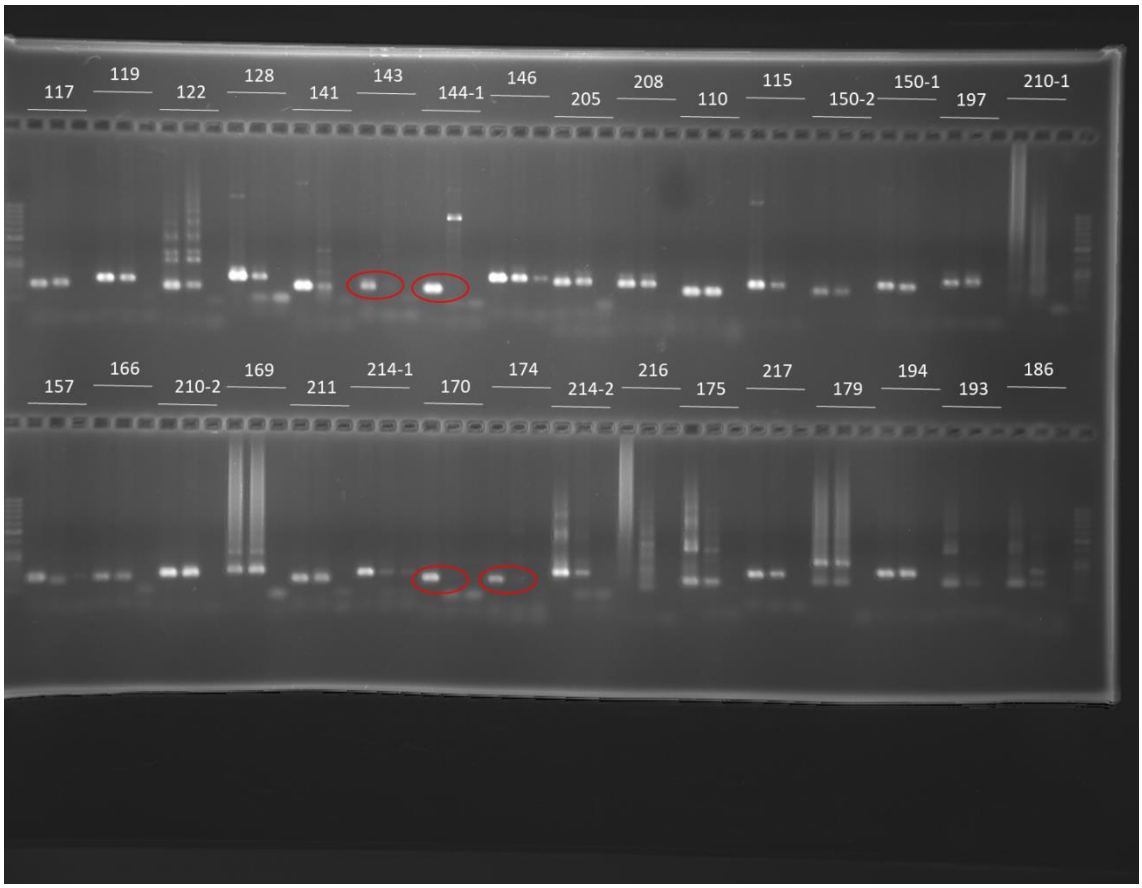
4.2.8 Charakteristika TAREAN anotací a identifikovaných satelitních sekvencí

U jednotlivých souborů bylo odhaleno 11 až 13 potencionálních tandemových repetic, což představuje přibližně 5 % ze všech anotovaných klastrů. Nejčastěji byly klastry anotovány jako satelity s nízkou spolehlivostí (53 až 83 % z počtu odhalených tandemových satelitů). Mezi typem vstupního souboru a počtem TAREAN anotací nebyl nalezen náznak korelace. Konkrétní počty jsou zobrazeny v Tab. 7. Na základě rozdílností v délkách identifikovaných sekvencí (Obr. 16), lze pozorovat silnou korelaci mezi délkou elementů a délkou vstupních čtení.

4.3 Navržené primery

Tab. 7 Počty *TAREAN* anotací u klastrů. Anotace vznikly aplikací algoritmu *RepeatExplorer2 clustering* na jednotlivé typy vstupních souborů

Typ vstupu	S duplikacemi			Bez duplikací		
Délka čtení	100	150	200	100	150	200
Unknown	230	227	211	224	230	230
Putative satellites (low confidence)	8	9	7	10	7	7
Putative LTR elements	1	2	3	1	2	2
Putative satellites (high confidence)	1	0	2	0	1	1
rDNA	1	1	1	1	1	1



Obr. 17 Výsledky gelové elektroforézy při využití navržených primerů k PCR amplifikaci. Každý primer byl využit na tři vzorky – s přítomností B chromozomů, s absencí B chromozomů a negativní kontrola. Červěně jsou označeny B specifické markery.

Na základě sekvencí odvozených od B specifických klastrů bylo navrženo 32 sad primerů. Jednotlivé primery byly využity při PCR amplifikaci DNA vzorků s přítomností a absencí B chromozomů. Namnožené fragmenty DNA byly separovány gelovou elektroforézou a porovnány. V případě čtyř sad primerů se podařilo získat fragmenty specifické pro B chrommozom (Obr. 17), tyto primery lze využít při charakterizaci genotypu *Sorghum purpureosericeum.*

5 ZÁVĚR

V této diplomové práci byly analyzovány repetitivní sekvence DNA v genomu rostliny *Sorghum purpureosericeum*, za účelem nalezení repetic specifických pro B chromozomy, které se v genomu mohou nacházet, díky kterým by se dal identifikovat genotyp rostliny.

Genomická data ze dvou vzorků (s přítomností a absencí B chromozomů) byla sérií úprav modifikovaná do podoby vstupních souborů pro software *RepeatExplorer*, který na základě klastrové analýzy identifikuje a anotuje repetitivní sekvence v genomu. Tímto způsobem byly analyzovány repetitivní sekvence ve vzorcích lišící se nejen délkou čtení, ale i přítomností duplicitních čtení. Ke zpracování výsledků softwaru *RepeatExplorer* byl vyvinut software v programovacím jazyce Python, jehož hlavní funkcí je vizualizace anotovaných klastrů a čtení a extrakce informací o B specifických klastrech z výstupního archivu algoritmu *RepeatExplorer2 clustering*.

V závislosti na typu vstupních dat bylo identifikováno 10-16 B-specifických klastrů (obsahovaly více jak 80 % B čtení) a na základě jejich sekvence bylo využitím softwaru *Primer3* navrženo 32 sad primerů, z nichž 4 byly identifikovány jako B specifické, a proto je lze využít pro genotypizaci vzorku pomocí PCR. Průměrná délka B-specifických se u všech typů vstupních souborů pohybovala okolo 500 bp.

Rovněž byla na základě anotovaných čtení a klastrů provedena kvantitativní analýza repetitivních sekvencí v genomu *Sorghum Purpureosericeum*. Bylo odhaleno, že mezi repetitivními elementy DNA převažují transpozibilní elementy, nejvíce rodiny retrotranspozónů *Athilla, Tekay* a *Retand* z řádu *Ty3/Gypsy*. V souladu s nejčetnějšími elementy byly i nejčastěji anotované proteinové domény transpozonů, konkrétně domény *INT, GAG, RT*, či *RH* z řádu *Ty3/Gypsy*. Mimo to bylo 8-10 klastrů z každého vstupního souboru anotováno jako satelitní sekvence a jednotky klastrů jako rDNA, či ptDNA.

Součástí běhu softwaru *RepeatExplorer2* byla také *TAREAN* anotace klastrů, tedy identifikace tandemově repetitivních sekvencí. Takto bylo anotováno přibližně 10 satelitních sekvencí (většina s nízkou spolehlivostí), tyto počty odpovídají stejně anotovaným klastrům. V rámci tandemových repetic byla odhalena silná korelace mezi délkou vstupních čtení a délkou nalezených repetic.

Takto popsaná kompozice repetitivní DNA v genomu *Sorghum purpureosericeum* je velmi podobná příbuznému *Sorghum bicolor*, což naznačuje blízkou příbuznost těchto druhů a vhodnost zvoleného postupu pro danou analýzu (Paterson *et al*., 2009).

6 LITERATURA

Adega F., Chaves R., Guedes-Pinto H. (2007): Chromosomal evolution and phylogenetic analyses in Tayassu pecari and Pecari tajacu (Tayassuidae): tales from constitutive heterochromatin. *Journal of genetics* **86**(1), 19–26.

Afgan E., Baker D., Batut B., van den Beek M., Bouvier D., Cech M., Chilton J., Clements D., Coraor N., Grüning B. A., Guerler A., Hillman-Jackson J., Hiltemann S., Jalili V., Rasche H., Soranzo N., Goecks J., Taylor J., Nekrutenko A., Blankenberg D. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic acids research*, **46**(1), 537–544.

Aldrich J.C., Ferree P.M. (2017): Genome silencing and elimination: Insights from a "selfish" B chromosome. *Front. Genet.* **8**, 50.

Alfenito, M. R., & Birchler, J. A. (1993): Molecular characterization of a maize B chromosome centric sequence. *Genetics* **135**(2), 589–597.

Arensburger, P., Hice, R.H., Wright, J.A. et al. (2011): The mosquito Aedes aegypti has a large genome size and high transposable element load but contains a low proportion of transposon-specific piRNAs. *BMC Genomics* **12**(606).

Banaei-Moghaddam A. M., Meier K., Karimi-Ashtiyani R., Houben A. (2013): Formation and expression of pseudogenes on the B chromosome of rye. *The Plant cell* **25**(7), 2536–2544.

Barski A., Cuddapah S., Cui K., Roh T. Y., Schones D. E., Wang Z., Wei G., Chepelev I., Zhao K. (2007): High-resolution profiling of histone methylations in the human genome. *Cell* **129**(4), 823–837.

Bennetzen J. L., Wang H. (2014): The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annual review of plant biology* **65**, 505–530.

Beukeboom L. (1994): Bewildering Bs: an impression of the 1st B-Chromosome Conference. *Heredity* **73**, 328–336.

Biscotti M. A., Olmo E., Heslop-Harrison J. S. (2015): Repetitive DNA in eukaryotic genomes. *Chromosomeresearch* **23**, 415-420.

Bolger A. M., Lohse M., Usadel B. (2014): Trimmomatic: A flexible trimer for Illumina Sequence Data. *Bioinformatics* **30**(15), 2114-2110.

Britten R. J., Graham D. E., Neufeld B. R. (1974): Analysis of repeating DNA sequences by reassociation. *Methods in enzymology* **29**, 363–418.

Brown T. A. (2002): *Genomes*. 2nd edition. Oxford: Wiley-Liss;. Chapter 2, Genome Anatomies. Available from: https://www.ncbi.nlm.nih.gov/books/NBK21120/#

Camacho J. P., Sharbel T. F., Beukeboom L. W. (2000): B-chromosome evolution. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **355**(1394), 163–178.

Clauset A., Newman M. E. J, Moore C. (2004): Finding community structure in very large networks. *Physical Review E* **70**(6), 066111.

Costa F. F. (2008): Non-coding RNAs, epigenetics and complexity. *Gene* **410**(1), 9-17.

Csardi G., Nepusz T. (2006): The igraph software package for complex network research. *InterJournal*, *Complex Systems* **1695**, https://igraph.org.

del Bosque M. E., Navajas-Pérez R., Panero J. L., Fernández-González A., Garrido-Ramos M. A. (2011): A satellite DNA evolutionary analysis in the North American endemic dioecious plant Rumex hastatulus (Polygonaceae). *Genome* **54**(4), 253–260.

Dessau, R. B., Pipper, C. B. (2008): R--en programpakke til statistisk databehandling og grafik ["R"--project for statistical computing]. *Ugeskrift for laeger* **170**(5), 328–330.

Dewannieux M., Esnault C., Heidmann T. (2003): LINE-mediated retrotransposition of marked Alu sequences. *Nature genetics* **35**(1), 41–48.

Donald T. M., Houben A., Leach C. R., Timmis J. N. (1997): Ribosomal RNA genes specific to the B chromosomes in Brachycome dichromosomatica are not transcribed in leaf tissue. *Genome* **40**(5), 674–681.

Du C., Swigonová Z., Messing J. (2006): Retrotranspositions in orthologous regions of closely related grass species. *BMC evolutionary biology* **6**, 62.

Endo, T. R., Nasuda, S., Jones, N., Dou, Q., Akahori, A., Wakimoto, M., *et al*. (2008): Dissection of rye B chromosomes, and nondisjunction properties of the dissected segments in a common wheat background. *Genes Genet. Syst*. **83**, 23–30.

Feschotte C., Pritham E. J. (2007): DNA transposons and the evolution of eukaryotic genomes. *Annual review of genetics* **41**, 331–368.

Finnegan D. J. (1989): Eukaryotic transposable elements and genome evolution. *Trends in genetics: TIG* **5**(4), 103–107.

Fuerst J. A. (2010): Beyond Prokaryotes and Eukaryotes: Planctomycetes and Cell Organization. *Nature Education* **3**(9), 44.

Gall J. G. (2016): The origin of in situ hybridization – A personal history. *Methods (San Diego, Calif.), 98*, 4–9.

Garrido-Ramos M. A. (2015): Satellite DNA in Plants: More than Just Rubbish. *Cytogenetic and genome research* **146**(2), 153–170.

Garrido-Ramos M. A. (2017): Satellite DNA: An Evolving Topic. *Genes* **8**(9), 230.

Garrido-Ramos M. A., Jamilena M., Lozano R., Cárdenas S., Ruiz Rejón C., Ruiz Rejón M. (1995): Phylogenetic Relationships of the *Sparidae* Family (*Pisces, Perciformes*) Inferred from Satellite-DNA *Hereditas* **122**(1), 1-6.

Girvan M., Newman M. E. J. (2002): Community structure in social and biological networks. *PNAS* **99**(12), 7821-7826.

Grabundzija, I., Messing, S. A., Thomas, J., Cosby, R. L., Bilic, I., Miskey, C., Gogol-Döring, A., Kapitonov, V., Diem, T., Dalda, A., Jurka, J., Pritham, E. J., Dyda, F., Izsvák, Z., & Ivics, Z. (2016): A Helitron transposon reconstructed from bats reveals a novel mechanism of genome shuffling in eukaryotes. *Nature communications* **7**, 10716.

Guio L., González J. (2019): New Insights on the Evolution of Genome Content: Population Dynamics of Transposable Elements in Flies and Humans. *Evolutionary Genomics. Methods in Molecular Biology*, vol 1910. Humana, New York, NY.

Houben A. (2017): B Chromosomes – A Matter of Chromosome Drive. *Frontiers in Plant Science* **8**, 210.

Houben A., Leach C. R., Verlin D., Rofe R., Timmis J. N. (1997): A repetitive DNA sequence common to the different B chromosomes of the genus Brachycome. *Chromosoma* **106**(8), 513–519.

Houben A., Verlin D., Leach C. R., Timmis J. N. (2001): The genomic complexity of micro B chromosomes of Brachycome dichromosomatica. *Chromosoma* **110**(7), 451–459.

Chakraborty M., Baldwin-Brown J. G., Long A. D., Emerson J. J. (2016): Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic acids research* **44**(19), e147.

Charlesworth B., Sniegowski P., Stephan W. (1994): The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature* **371**(6494), 215–220.

Cheng Y. M., & Lin B. Y. (2004): Molecular organization of large fragments in the maize B chromosome: indication of a novel repeat. *Genetics* **166**(4), 1947–1961.

Cheng Y. M., Lin B. Y. (2003): Cloning and characterization of maize B chromosome sequences derived from microdissection. *Genetics* **164**(1), 299–310.

Cheng Z., Stupar R., Gu M. *et al.* (2001): A tandemly repeated DNA sequence is associated with both knob-like heterochromatin and a highly decondensed structure in the meiotic pachytene chromosomes of rice. *Chromosoma* **110,** 24–31.

Cheng, Z., Stupar, R. M., Gu, M., Jiang, J. (2001): A tandemly repeated DNA sequence is associated with both knob-like heterochromatin and a highly decondensed structure in the meiotic pachytene chromosomes of rice. *Chromosoma* **110**(1), 24–31.

Jain M., Olsen H. E., Turner D. J., Stoddart D., Bulazel K. V., Paten B., Haussler D., Willard H. F., Akeson M., Miga K. H. (2018): Linear assembly of a human centromere on the Y chromosome. *Nature biotechnology*, **36**(4), 321–323.

Jiménez G., Manzanero S., & Puertas M. J. (2000): Relationship between pachytene synapsis, metaphase I associations, and transmission of 2B and 4B chromosomes in rye. *Genome* **43**(2), 232–239.

Jiménez M., Romera F., González-Sánchez M., Puertas M. J. (1997): Genetic control of the rate of transmission of rye B chromosomes. III. Male meiosis and gametogenesis. *Heredity* **78**, 636–644.

Jones R. N., Viegas W., Houben A. (2008): A century of B chromosomes in plants: so what? *Annals of botany* **101**(6), 767–775.

Juergen B. (2009): The Fragmented Gene. *Annals of the New York Academy of Sciences* **1178**, 186-93.

Kapitonov V. V., Jurka J. (2006): Self-synthesizing DNA transposons in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America* **103**(12), 4540–4545.

Khost D. E., Eickbush D. G., Larracuente A. M. (2017): Single-molecule sequencing resolves the detailed structure of complex satellite DNA loci in *Drosophila melanogaster*. *Genome research* **27**(5), 709–721.

Kit S. (1961): Equilibrium sedimentation in density gradients of DNA preparations from animal tissues. *Journal of molecular biology* **3**, 711–716.

Kit S. (1961): Equilibrium sedimentation in density gradients of DNA preparations from animal tissues. *Journal of molecular biology* **3**, 711–716.

Klemme S., Banaei-Moghaddam A. M., Macas J., Wicker T., Novák P., Houben A. (2013): High-copy sequences reveal distinct evolution of the rye B chromosome. *The New phytologist* **199**(2), 550–558.

Klemme S., Banaei-Moghaddam A.M., Macas J., Wicker T., Novák P., Houben A. (2013): High-copy sequences reveal distinct evolution of the rye B-chromosome. *New Phytol.* **199**, 550–558.

Kofler, R., Nolte, V., & Schlötterer, C. (2015): Tempo and Mode of Transposable Element Activity in Drosophila. *PLoS genetics* **11**(7), e1005406.

Koressaar T., Remm M. (2007): Enhancements and modifications ofprimer design program Primer3. *Bioinformatics* **23**(10), 1289-1291.

Krupovic M., Koonin E. V. (2016): Self-synthesizing transposons: unexpected key players in the evolution of viruses and defense systems. *Current opinion in microbiology* **31**, 25–33.

Lai J., Li Y., Messing J., Dooner, H. K. (2005): Gene movement by Helitron transposons contributes to the haplotype variability of maize. *Proceedings of the National Academy of Sciences of the United States of America* **102**(25), 9068–9073.

Lam E. T., Hastie A., Lin C., Ehrlich D., Das S. K., Austin M. D., Deshpande P., Cao H., Nagarajan N., Xiao M., Kwok P. Y. (2012): Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nature biotechnology* **30**(8), 771–776.

Lamb J. C., Riddle N. C., Cheng Y. M., Theuri J., Birchler J. A. (2007): Localization and transcription of a retrotransposon-derived element on the maize B chromosome. *Chromosome research: an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, **15**(3), 383–398.

Lamb, J. C., Han, F., Auger, D. L., and Birchler, J. A. (2006): A trans-acting factor required for non-disjunction of the B chromosome is located distal to the TB-4Lb breakpoint on the B chromosome. *Maize Genet. Coop. News Lett*. **80**, 51–54.

Langdon T., Seago C., Jones R. N., Ougham H., Thomas H., Forster J. W., Jenkins G. (2000): De novo evolution of satellite DNA on the rye B chromosome. *Genetics* **154**(2), 869–884.

Lawrence M., Wickham H., Cook D., Hofmann H., Swayne D. F. (2009): Extending the GGobi pipeline from R. *Comput Stat* **24,** 195–205.

Leach C. R., Donald T. M., Franks T. K., Spiniello S. S., Hanrahan C. F., Timmis J. N. (1995): Organisation and origin of a B chromosome centromeric sequence from Brachycome dichromosomatica. *Chromosoma* **103**(10), 708–714.

Leach C. R., Houben A., Field B., Pistrick K., Demidov D., Timmis J. N. (2005): Molecular evidence for transcription of genes on a B chromosome in Crepis capillaris. *Genetics* **171**(1), 269–278.

Lee S. I., & Kim N. S. (2014): Transposable elements and genome size variations in plants. *Genomics & informatics* **12**(3), 87–97.

Li H., Durbin R. (2010): Fast and accurate long-read alignment with Burrows-Wheeler Transform. *Bioinformatics* **26**(5), 589-95.

Lodish H., Berk A., Kaiser Ch. A., Krieger M., Bretscher A., Ploegh H., Amon A., Martin K. C. (2016): *Molecular Cell Biology*. 8th ed., W. H. Freeman, New York, 1280 stran.

Longley A. E. J (1927): Supernumerary chromosomes in Zea mays. *Agric Res* **35**, 769–784.

López-Flores I., Garrido-Ramos M. A. (2012): The repetitive DNA content of eukaryotic genomes. *Genome dynamics*, **7**, 1–28.

López-León M. D., Neves N., Schwarzacher T., Heslop-Harrison J. S., Hewitt G. M., Camacho J. P. (1994): Possible origin of a B chromosome deduced from its DNA composition using double FISH technique. *Chromosome research: an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology* **2**(2), 87–92.

Lorenzi H. A., Puiu D., Miller J. R., Brinkac L. M., Amedeo P., Hall N., Caler E. V. (2010): New assembly, reannotation and analysis of the Entamoeba histolytica genome reveal new genomic features and protein content information. *PLoS neglected tropical diseases* **4**(6).

Lower S. S., McGurk M. P., Clark A. G., Barbash D. A. (2018): Satellite DNA evolution: old ideas, new approaches. *Current opinion in genetics & development* **49**, 70–78.

Macas, J., Mészáros, T., Nouzová, M. (2002): PlantSat: a specialized database for plant satellite repeats. *Bioinformatics* **18**(1), 28–35.

Macas, J., Neumann, P., Navrátilová, A. (2007): Repetitive DNA in the pea (Pisum sativum L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and Medicago truncatula. *BMC genomics*, **8**, 427.

Maluszynska J., Schweizer D. (1989): Ribosomal RNA genes in B chromosomes of Crepis capillaris detected by non-radioactive in situ hybridization. *Heredity* **62**, 59–65.

Marques A., Klemme S., Houben A. (2018): Evolution of Plant B Chromosome Enriched Sequences. *Genes* **9**(10), 515.

Martis M. M., Klemme S., Banaei-Moghaddam A. M., Blattner F. R., Macas J., Schmutzer T., Scholz U., Gundlach H., Wicker T., Šimková H., Novák P., Neumann P., Kubaláková M., Bauer E., Haseneyer G., Fuchs J., Doležel J., Stein N., Mayer K. F., Houben A. (2012): Selfish supernumerary chromosome reveals its origin as a mosaic of host genome and organellar sequences. *Proceedings of the National Academy of Sciences of the United States of America* **109**(33), 13343–13346.

Maxwell P. H. (2020): Diverse transposable element landscapes in pathogenic and nonpathogenic yeast models: the value of a comparative perspective. *Mobile DNA* **11**(16).

Mehrotra S., Goel S., Sharma S. *et al.* (2013): Sequence Analysis of *Kpn*I Repeat Sequences to Revisit the Phylogeny of the Genus *Carthamus* L. *Appl Biochem Biotechnol* **169,** 1109–1125.

Mehrotra, S., Goyal, V. (2014): Repetitive sequences in plant nuclear DNA: types, distribution, evolution and function. *Genomics, proteomics & bioinformatics* **12**(4), 164–171.

Melo, E. S., & Wallau, G. L. (2020). Mosquito genomes are frequently invaded by transposable elements through horizontal transfer. *PLoS genetics*, **16**(11).

Melters D. P., Bradnam K. R., Young H. A. *et al.* (2013): Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol* **14,** R10.

Melters, D. P., Bradnam, K. R., Young, H. A., Telis, N., May, M. R., Ruby, J. G., Sebra, R., Peluso, P., Eid, J., Rank, D., Garcia, J. F., DeRisi, J. L., Smith, T., Tobias, C., Ross-Ibarra, J., Korf, I., & Chan, S. W. (2013): Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome biology* **14**(1), R10.

Nassif N., Penney J., Pal S., Engels W. R., Gloor G. B. (1994): Efficient copying of nonhomologous sequences from ectopic sites via P-element-induced gap repair. *Molecular and cellular biology* **14**(3), 1613–1625.

Newman M. E. J. (2006): Modularity and community structure in networks. *PNAS* **103**(23), 8577-8582.

Newman M. E. J., Girvan M. (2004): Finding and evaluating community structure in networks. *Physical Review E* **69**(2), 026113.

Noboru S. (1961): Variation and heterogeneity of base composition of deoxyribonucleic acids: A compilation of old and new data. *Journal of Molecular Biology* **3**(1), 31-40.

Novák P., Ávila Robledillo L., Koblížková A., Vrbová I., Neumann P., Macas J. (2017): TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic acids research* **45**(12), e111.

Novák P., Ávila Robledillo L., Koblížková A., Vrbová I., Neumann P., Macas J. (2017): TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic acids research* **45**(12), e111.

Novák, P., Neumann, P., Macas, J. (2010): Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics* **11,** 378.

Nur U., Werren J. H., Eickbush D. G., Burke W. D., Eickbush T. H. (1988): A "selfish" B chromosome that enhances its transmission by eliminating the paternal genome. *Science* **240**, 512–514.

Ohtsubo, H., & Ohtsubo, E. (1994): Involvement of transposition in dispersion of tandem repeat sequences (TrsA) in rice genomes. *Molecular & general genetics:MGG* **245**(4), 449–455.

Pardue M. L., Gall J. G. (1970): Chromosomal localization of mouse satellite DNA. *Science (New York, N.Y.), 168*(3937), 1356–1358.

Paterson A. H., Bowers J. E., Bruggmann R., Dubchak I., Grimwood J., Gundlach H., Haberer G., Hellsten U., Mitros T., Poliakov A., Schmutz J., Spannagl M., Tang H., Wang X., Wicker T., Bharti A. K., Chapman J., Feltus F. A., Gowik U., Grigoriev I. V., *et al*. (2009): The Sorghum bicolor genome and the diversification of grasses. *Nature*, **457**(7229), 551–556.

Perfectti F., & Werren J. H. (2001): The interspecific origin of B chromosomes: experimental evidence. *Evolution; international journal of organic evolution* **55**(5), 1069–1073.

Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., Lee, Y., White, J., Cheung, F., Parvizi, B., Tsai, J., Quackenbush, J. (2003): TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics (Oxford, England)*, **19**(5), 651–652.

Plohl M., Luchetti A., Mestrović N., Mantovani B. (2008). Satellite DNAs between selfishness and functionality: structure, genomics and evolution of tandem repeats in centromeric (hetero)chromatin. *Gene* **409**(1-2), 72–82.

Plohl M., Meštrović N., Mravinac B. (2012): Satellite DNA evolution. *Genome dynamics* **7**, 126–152.

Quesada del Bosque M. E., López-Flores I., Suárez-Santiago V. N., Garrido-Ramos M. A. (2013): Differential spreading of HinfI satellite DNA variants during radiation in Centaureinae. *Annals of botany* **112**(9), 1793–1802.

R Core Team (2020): R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/.

Robinson T. J., Thorvaldsdóttir H., Wincler W., Guttman M., Lander E. S., Getz G., Mesirov J. P. (2011): Integrative Genomics Viewer. *Nature Biotechnology* **29**(1), 24-26.

Rowold D. J., Herrera R. J. (2000): Alu elements and the human genome. *Genetica* **108**(1), 57–72.

Ruban A., Schmutzer T., Scholz U., Houben A. (2017): How Next-Generation Sequencing Has Aided Our Understanding of th Sequence Composition and Origin of B Chromosomes. *Genes* **294**(8).

Ruiz-Ruano F. J., López-León M. D., Cabrero J., Camacho J. (2016): High-throughput analysis of the satellitome illuminates satellite DNA evolution. *Scientific reports* **6**, 28333.

Sevim, V., Bashir, A., Chin, C. S., & Miga, K. H. (2016). Alpha-CENTAURI: assessing novel centromeric repeat sequence variation with long read sequencing. *Bioinformatics (Oxford, England)* **32**(13), 1921–1924.

Shirasu K., Schulman A. H., Lahaye T., Schulze-Lefert P. (2000): A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. *Genome research* **10**(7), 908–915.

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T. A., Minx, P., Reily, A. D., Courtney, L., Kruchowski, S. S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S. M., … Wilson, R. K. (2009): The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**(5956), 1112–1115.

Sijen T., Plasterk R.H. (2003): Transposon silencing in the *Caenorhabditis elegans* germ line by natural RNAi. *Nature* **426**, 310–314.
Singer M. F. (1982): Highly repeated sequences in mammalian genomes. *International review of cytology* **76**, 67–112.
Spaller, T., Kling, E., Glöckner, G., Hillmann, F., & Winckler, T. (2016): Convergent evolution of tRNA gene targeting preferences in compact genomes. *Mobile DNA* **7**(1), 17.
Stark E. A., Connerton I., Bennett S. T., Barnes S. R., Parker J. S., Forster J. W. (1996): Molecular analysis of the structure of the maize B-chromosome. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology* **4**(1), 15–23.
Sultana, T., Zamborlini, A., Cristofari, G., & Lesage, P. (2017): Integration site selection by retroviruses and transposable elements in eukaryotes. *Nature reviews. Genetics* **18**(5), 292–308.
Swayne D. F., Lang D. T., Buja A., Cook D. (2003): GGobi: evolving from XGobi into an extensible framework for interactive data visualization. *Comput Stat Data An*. **43**(4), 423–444.
Twell, D. (2010): Male gametogenesis and germline specification in flowering plants. *Sex. Plant Reprod*. **24**, 149–160.
Ueng P. P., Hang A., Tsang H., Vega J. M., Wang L., Burton C. S., He F. T., Liu B. (2000): Molecular analyses of a repetitive DNA sequence in wheat (Triticum aestivum L.). *Genome*. **43**(3), 556-563.
Vogt P. (1990): Potential genetic functions of tandem repeated DNA sequence blocks in the human genome are based on a highly conserved "chromatin folding code". *Hum Genet* **84,** 301–336.
Vývojový tým pandas (2021): https://zenodo.org/record/4572994#.YEuKk9wo99A
Walsh J. B. (1987). Persistence of tandem arrays: implications for satellite and simple-sequence DNAs. *Genetics* **115**(3), 553–567.
Waring M., & Britten R. J. (1966): Nucleotide sequence repetition: a rapidly reassociating fraction of mouse DNA. *Science (New York, N.Y.)*, **154**(3750), 791–794.
Wessler S. R., Bureau T. E., White S. E. (1995): LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. *Current opinion in genetics & development* **5**(6), 814–821.
Wicker T., Sabot F., Hua-Van A., Bennetzen J. L., Capy P., Chalhoub B., Flavell A., Leroy P., Morgante M., Panaud O., Paux E., SanMiguel P., Schulman A. H. (2007): A unified classification system for eukaryotic transposable elements. *Nature reviews. Genetics* **8**(12).
Willard H. F. (1985): Chromosome-specific organization of human alpha satellite DNA. *American journal of human genetics* **37**(3), 524–532.
Woehle, C., Kusdian, G., Radine, C. et al. (2014): The parasite Trichomonas vaginalis expresses thousands of pseudogenes and long non-coding RNAs independently from functional neighbouring genes. *BMC Genomics* **15**(906).
Xu H., Luo X., Qian J., Pang X., SongJ., Qian G., Chen J., Chen S. (2012): FastUniq: A Fast *De Novo* Duplicates Removal Tool for Paired Short Reads. *PLoS One* **7**(12):e52249.
Zuccolo A., Sebastian, A., Talag J., Yu Y., Kim H., Collura K., Kudrna D., Wing R. A. (2007): Transposable element distribution, abundance and role in genome size variation in the genus Oryza. *BMC evolutionary biology*, **7**(152).

7 SEZNAM POUŽITÝCH ZKRATEK

bp – pár nukleotidových bází

kb – kilobáze

Mb – megabáze

Gb – gigabáze

8 PŘÍLOHY

8.1 Tabulka se srovnáním počtů a typů anotovaných proteinových domén transpozibilních elementů

Typ vstupu	S duplikacemi			Bez duplikací		
Délka čtení	100	150	200	100	150	200
Ty3-INT	20851	21588	21528	21911	21977	22180
Ty3-RT	12581	12922	13953	12984	13659	14130
Ty3-GAG	10910	12438	13792	11073	12964	14114
Ty3-RH	9973	10669	11338	10311	10971	11795
Ty3-PROT	5592	7513	9197	5669	7564	9175
Ty1-RT	4922	5474	6064	4762	5703	6187
CACTA-TPase	4374	4504	4747	4596	4626	4909
Ty1-INT	3322	3756	4121	3515	3696	4253
Ty3-CHD	2576	2870	3449	2551	3120	3440
Ty3-aRH	2470	2692	2622	2655	2628	2760
Ty1-RH	2429	2974	3697	2552	3276	3687
Ty1-GAG	1291	1572	1930	1360	1673	1959
Ty1-PROT	1103	1413	1761	1179	1432	1859
LINE-RT	499	560	723	666	669	699
LINE-ENDO	314	303	383	265	260	353
MuDR-TPase	127	169	210	34	131	151
Ty3-CHDCR	71	96	123	63	108	120
hAT-TPase	27	41	50	30	33	63
Harbinger-TPase	0	4	8	4	8	8
PARA-PROT	0	3	1	1	3	5
Helitron-HEL1	0	1	52	0	1	7
Helitron-HEL2	0	0	75	0	1	93
PARA-RH	0	0	1	0	1	0
Mariner-TPase	0	0	0	0	0	1
PARA-RT	0	0	0	0	0	1

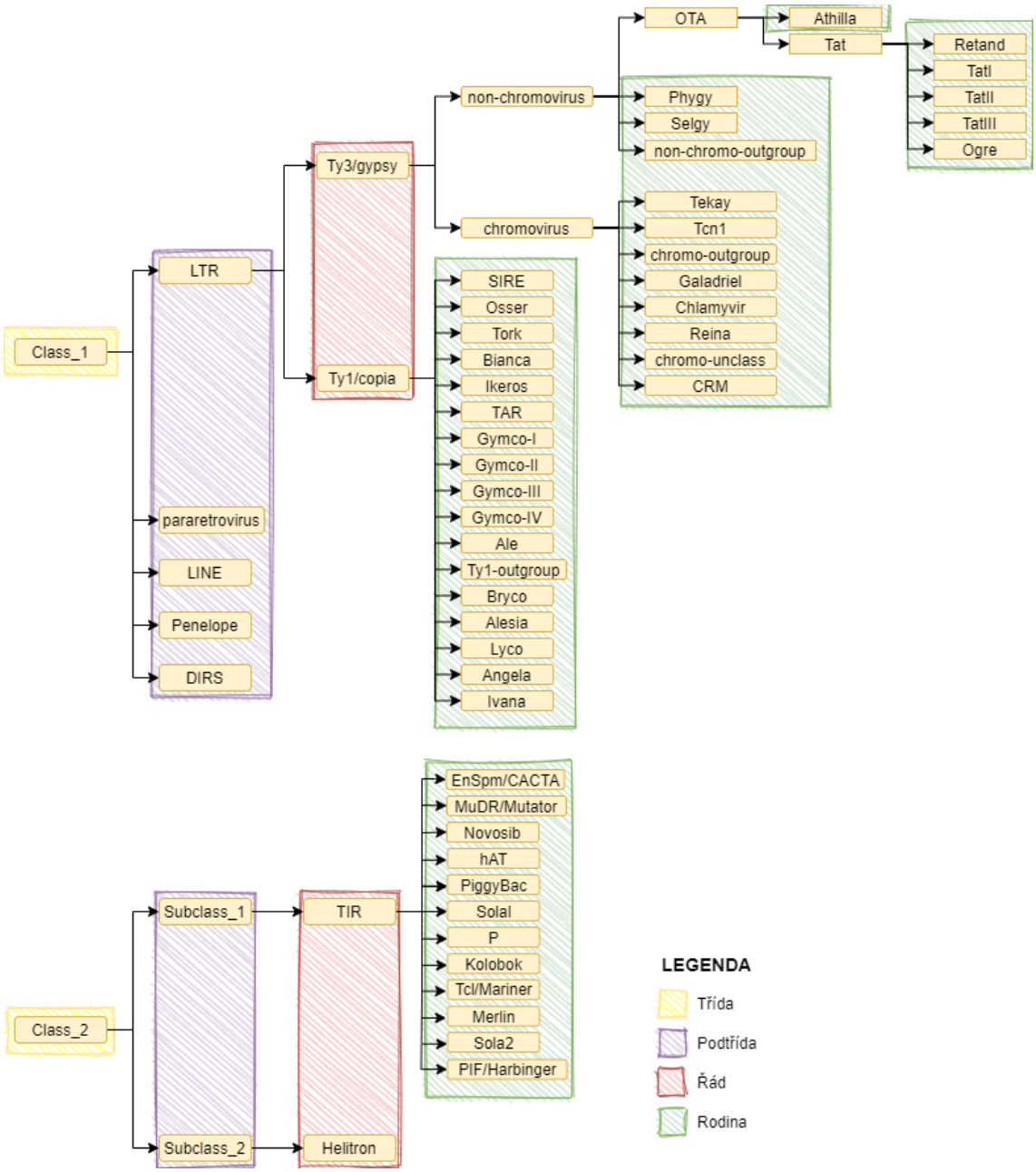
8.2 Tabulka se srováním počtů a typů rodin anotovaných transpozibilních elementů

Typ vstupu	S duplikacemi			Bez duplikací		
Délka čtení	100	150	200	100	150	200
Tekay	29931	31913	34166	30951	33102	34890
Retand	15618	16867	17497	16365	17165	18109
Athila	13935	16118	17994	14438	16706	18093
SIRE	10733	11907	13181	10952	12332	13433
Ogre	5228	5497	5707	5160	5616	5922
EnSpm_CACTA	4374	4504	4747	4596	4626	4909
TAR	841	940	1154	947	1008	1201
LINE	813	863	1106	931	929	1052
Ikeros	651	809	764	615	860	894
Bianca	481	581	744	420	640	752
Ivana	314	644	801	337	645	826
CRM	271	343	557	267	351	616
MuDR_Mutator	127	169	210	34	131	151
Tork	46	125	423	0	132	281
hAT	27	41	50	30	33	63
Reina	24	22	45	16	31	48
Tcn1	11	12	16	13	6	10
TatII	2	0	0	4	1	0
Ale	1	175	360	96	147	415
Galadriel	1	12	14	1	9	10
Phygy	1	0	0	1	0	1
TatIII	1	2	0	0	0	1
chromo-unclass	1	0	5	1	2	4
Alesia	0	0	0	1	1	0
PIF_Harbinger	0	4	8	4	8	8
pararetrovirus	0	3	2	1	4	6
Angela	0	4	141	0	13	140
Helitron	0	1	127	0	2	100
Osser	0	4	2	0	0	1
Selgy	0	1	0	0	0	3
non-chromo-outgroup	0	1	0	0	0	3
Chlamyvir	0	0	1	0	2	4
Gymco-II	0	0	2	0	1	2
Ty1-outgroup	0	0	0	0	1	0
Bryco	0	0	1	0	0	0
Tc1_Mariner	0	0	0	0	0	1

8.3 Tabulka se srovnáním počtů a typů anotací klastrů

Typ vstupu	S duplicitami			Bez duplicit		
Délka čtení	100	150	200	100	150	200
All	77	74	69	74	70	65
Athila	35	34	32	36	35	30
Tekay	35	37	36	36	38	35
Retand	23	20	17	19	20	15
SIRE	15	15	12	16	17	13
EnSpm_CACTA	14	11	10	13	14	11
satellite	10	8	8	10	8	7
LTR	7	6	5	8	9	8
Ogre	7	6	5	8	5	4
Bianca	4	3	2	3	3	2
CRM	3	2	5	1	3	5
MuDR_Mutator	3	4	3	2	2	3
LINE	2	2	3	3	2	5
25S_rDNA	1	1	1	1	1	1
45S_rDNA	1	1	1	1	1	1
5S_rDNA	1	1	1	1	1	1
Ikeros	1	2	1	1	3	1
TAR	1	2	2	1	2	3
Tork	1	1	3	0	1	2
Ale	0	2	2	1	1	4
Tat	0	0	0	1	0	0
Class_I	0	1	0	0	0	0
Ivana	0	1	2	0	1	2
mobile_element	0	3	0	0	0	2
plastid	0	1	1	0	2	2
repeat	0	1	1	0	1	1
Ty1_copia	0	0	0	0	1	0
Angela	0	0	1	0	0	3
Helitron	0	0	1	0	0	1

8.4 Strom klasifikace transpozibilních elementů využitý při anotacích softwarem *RepeatExplorer2*



8.5 Přílohy na přiloženém disku

1. Soubor *RE_process*.exe sloužící k instalaci implementovaného softwaru ke zpracování výsledků algoritmu *RepeatExplroer2 clustering.*

2. 6 výstupních složek softwaru vzniklých jeho aplikací na 6 typů vstupních dat, složky jsou pojmenovány jako *output_[dup|nodup]_<délka čtení>bp* a nacházejí se v adresáři *výsledky.*